IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 15, 2020, accepted May 30, 2020, date of publication June 4, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000179

A Taxonomy of Network Threats and the Effect of Current Datasets on Intrusion Detection Systems

HANAN HINDY[1], (Member, IEEE), DAVID BROSSET[2], ETHAN BAYNE[1], AMAR SEEAM[3], (Member, IEEE), CHRISTOS TACHTATZIS[4], (Senior Member, IEEE), ROBERT ATKINSON[4], (Senior Member, IEEE), AND XAVIER BELLEKENS[1,4], (Member, IEEE)

[1]Division of Cyber Security, Abertay University, Dundee DD1 1HG, U.K.
[2]Naval Academy Research Institute, 29240 Brest, France
[3]Department of Computer Science, Middlesex University, Uniciti 90203, Mauritius
[4]EEE Department, University of Strathclyde, Glasgow G1 1XQ, U.K.

Corresponding author: Hanan Hindy (hananhindy@ieee.org)

ABSTRACT As the world moves towards being increasingly dependent on computers and automation, building secure applications, systems and networks are some of the main challenges faced in the current decade. The number of threats that individuals and businesses face is rising exponentially due to the increasing complexity of networks and services of modern networks. To alleviate the impact of these threats, researchers have proposed numerous solutions for anomaly detection; however, current tools often fail to adapt to ever-changing architectures, associated threats and zero-day attacks. This manuscript aims to pinpoint research gaps and shortcomings of current datasets, their impact on building Network Intrusion Detection Systems (NIDS) and the growing number of sophisticated threats. To this end, this manuscript provides researchers with two key pieces of information; a survey of prominent datasets, analyzing their use and impact on the development of the past decade's Intrusion Detection Systems (IDS) and a taxonomy of network threats and associated tools to carry out these attacks. The manuscript highlights that current IDS research covers only 33.3% of our threat taxonomy. Current datasets demonstrate a clear lack of real-network threats, attack representation and include a large number of deprecated threats, which together limit the detection accuracy of current machine learning IDS approaches. The unique combination of the taxonomy and the analysis of the datasets provided in this manuscript aims to improve the creation of datasets and the collection of real-world data. As a result, this will improve the efficiency of the next generation IDS and reflect network threats more accurately within new datasets.

INDEX TERMS Anomaly detection, datasets, intrusion detection systems, network attacks, network security, security threats, survey, taxonomy.

I. INTRODUCTION

The world is becoming more dependent on connected devices, actuators and sensors, regulating the lives of millions of people. Furthermore, sensor data is expected to increase by around 13%, reaching 35% of overall data communication in 2020, reaching a peak of 50 billion connected devices and an increased monthly Internet traffic volume reaching 30 GB on average per capita compared to around 10 GB in 2016 [1]. While each device in an Internet of Things (IoT) system exchanges data, associated services often provide interfaces to interact with the collected data, often increasing the attack surface. Therefore, it is crucial to build robust tools to defend networks against security threats in modern IoT networks. Current detection tools are often based on outdated datasets that do not reflect the reality of recent/modern network attacks, rendering Intrusion Detection Systems (IDS) ineffective against new threats and zero-days. To the best knowledge of the authors, there are currently no manuscripts that analyze the shortcomings of available networking datasets, nor provide a taxonomy of the current network threats and the associated tools used to carry out these attacks.

The contributions of this research are threefold:

- An evaluation of the limitations of the available network-based datasets and their impact on the development of IDSs

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaofan He.

- A review of the last decade's research on NIDS
- A Threat taxonomy is presented, and categorized by:
  - The Threat Sources
  - The Open Systems Interconnection (OSI) Layer
  - Active or Passive modes

The evaluation of current network-based datasets provides researchers with an insight of the shortcomings of the datasets presented when used for training against real-world threats. A threat taxonomy is derived from the datasets and current real-world networking threats. This taxonomy serves two purposes— firstly, it strengthens our argument on the shortcomings of currently available datasets, but most importantly, it provides researchers with the ability to identify threats and tools underrepresented in currently available datasets. To facilitate this endeavor, we further map the current threats with their associated tools, which in turn can be used by research to create new datasets.

The rest of the paper is organized as follows; Section II depicts the main differences between IDSs, the metrics to consider for their evaluation and the role of feature selection in building IDSs. Section III reviews IDSs of the past decade and their individual contributions are assessed. This section also evaluates the drawbacks and limitations of the available datasets. Section IV provides the threat taxonomy. Section V summarizes the challenges presented in this work and provides recommendations. Finally, the paper is concluded in Section VI.

II. BACKGROUND

A. INTRUSION DETECTION SYSTEMS

IDSs are defined as systems built to monitor and analyze network traffic and/or systems to detect anomalies, intrusions or privacy violations. When an intrusion is detected, an IDS is expected to (a) log the information related to the intrusion, (b) trigger alerts and (c) take mitigation and corrective actions [2].

IDS can either be Host Intrusion Detection System (HIDS) or Network Intrusion Detection System (NIDS). HIDS is responsible for monitoring a system internally, having access to log files, users' activities, etc. While NIDS analyses incoming and outgoing communication between network nodes.

IDSs differ based on their detection method. Signature-based IDSs were the first to be developed. Accurate signatures are built from prior detected attacks. The main advantage of this method is the high accuracy of detecting known attacks. Signature-based IDS is, however, unable to detect zero-days, metamorphic and polymorphic threats [3]. The second method, Anomaly-based detection, depends on identifying patterns and comparing them to normal traffic patterns. This method requires the system to be trained prior to deployment. The accuracy of anomaly-based systems against zero-days, metamorphic and polymorphic threats is better when compared to signature-based IDS. However, the false positive rate of anomaly-based detection is often higher. It is important to mention that benign/normal traffic patterns alone are not sufficient to detect attacks. For this reason, the features used to represent network traffic play an essential role in traffic representation.

Intrusion detection (both signature-based and anomaly-based) can be done on a stateless (per packet) or stateful (per flow) basis. Most recent IDSs are stateful, as the flow provides "context", while packet analysis (stateless) does not provide this context. It is the responsibility of the researcher to decide which method is best suited for their application.

Anomaly-based IDS can be classified into subcategories based on the training method used. These categories are statistical, knowledge-based and Machine Learning (ML) based. Statistical includes univariate, multivariate and time series. Knowledge-based uses finite state machines and rules like case-based, N-based, expert systems and descriptor languages. Buczak and Guven [4] provide recommendations on choosing the ML/Deep Learning (DL) algorithms based on the problem intended to be solved. Algorithms include Artificial Neural Networks (ANN), clustering, Genetic Algorithms (GA), etc. Specification-based combines the strength of both signature and anomaly based to form a hybrid model.

Owezarski *et al.* [5] summarize the approaches to validate networking models, which applies to IDS, into four categories; mathematical models, simulation, emulation and real experiments. Each of these approaches has their own pros and cons as discussed by Behal and Kumar [6].

1) METRICS FOR IDS EVALUATION

In order for an IDS to be considered effective, high detection rate and low false positive rate are key aspects to consider. Multiple metrics could be used for an IDS evaluation. These metrics are discussed subsequently showing the significance and purpose of each. It is important to mention that depending only on detection rate as the only evaluation metric doesn't reflect an IDS performance.

Other important evaluation factors including the transparency and safety of the overall system, memory requirements, power consumption and throughput should be considered. Moreover, [7] adds to the aforementioned requirements, ease of use, interoperability, transparency and collaboration.

IDS accuracy can be defined in terms of:

- True Positive (TP): Number of intrusions correctly detected
- True Negative (TN): Number of non-intrusions correctly detected
- False Positive (FP): Number of non-intrusions incorrectly detected
- False Negative (FN): Number of intrusions incorrectly detected

Hodo *et al.* [8], Buse *et al.* [9] and Aminanto *et al.* [10] discuss the main metrics to consider for evaluation in their respective work. These include the overall accuracy, decision rates, precision, recall, F1 and Mcc.

Equation 1 provides the overall accuracy. It returns the probability that an item is correctly classified by the IDS.

$$OverallAccuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Equation 2 calculates the Sensitivity, Specificity, Fallout, and Miss Rate detection rates, respectively. Stefan Axelsson [7] stresses the fact that false positive rates (false alarms) highly limit the performance of an IDS due to the ''Base-rate fallacy problem''.

Detection Rates:

$$\begin{aligned} Sensitivity(aka\ Recall,\ True\ Positive\ Rate) &= \frac{TP}{TP + FN} \\ Specificity \\ (aka\ Selectivity,\ True\ Negative\ Rate) &= \frac{TN}{TN + FP} \\ Fallout(aka\ False\ Positive\ Rate) &= \frac{FP}{TN + FP} \\ Miss\ Rate(aka\ False\ Negative\ Rate) &= \frac{FN}{TP + FN} \end{aligned} \tag{2}$$

Equation 3 provides the percentage of positively classified incidents that are truly positive.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

To visualize the performance of an IDS, i.e. the trade-off between sensitivity (true positive rate) and fallout (true negative rate), AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) also known as Area Under the Receiver Operating Characteristics (AUROC) curve is used [11]–[13]

Equation 4 represents the harmonic mean of precision and recall. F1 is better suited to represent the performance of an IDS, specially when dealing with imbalanced classes.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Equation 5 provides Matthews correlation coefficient. It can only be used in binary IDS in which incidents are classified as either attack or normal.

$$Mcc = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Equation 6 addresses the problem of calculating accuracy in imbalanced datasets. Numerous datasets have a limited number of attacks' data compared to benign traffic, hence, the geometric mean of accuracy provides a more precise metric than overall accuracy measure [14]. Space-Time aware evaluation is introduced by Feargus Pendlebury *et al.* [15] to overcome both spatial and temporal biases. The authors introduced three constraints to be considered when splitting datasets.

$$\begin{aligned} gAcc &= \sqrt{a^{+ve}.a^{-ve}} \\ &= \sqrt{Sensitivity_{(TPR)}.Specificity_{(TNR)}} \end{aligned} \tag{6}$$

Additionally, CPU consumption, throughput and power consumption are important metrics for evaluating IDSs. Specifically, these metrics are important for IDSs running on different hardware or with specific settings such as high-speed networks, or on hardware with limited resources.

2) FEATURE SELECTION AND IDS

''Feature Learning'' [10] or ''Feature Engineering'' [16] plays an essential role in building any IDS in a way that chosen features highly affect the IDS performance. Features are obtained using one of three processes; construction, extraction and selection. Selection involves filter, wrapper and embedded techniques [17]. A classification of the features used in recent datasets is provided in [4].

Different features representations (i.e. abstractions) are used to address different areas of threat detection. Some of them could be considered naïve when they contain basic network information. Others are considered rich when they represent deeper details [16]. As highlighted by Rezaei and Liu [18], there are four main categories of networking features; time series, header, payload and statistical. Unlike header and payload features, time series and statistical ones are available for both encrypted and unencrypted traffic. The authors further discuss the shortcomings of current encrypted traffic classification research. Packet-based and flow-based features have been used for intrusion detection purposes. However, with the advancement of network encryption, packet-based features are rendered impractical for complex communication networks.

B. RELATED WORK

In the past decade, numerous IDSs were developed and evaluated against a range of published available datasets. Diverse review and comparative studies have been published tackling the design of IDS for various applications, as well as, the machine learning techniques used to build IDS, however, the dataset challenges are not discussed.

Current Network IDS surveys often focus on a single aspect of IDS evaluation. Buczak and Guven [4] focus on the different ML and DL algorithms used to build IDS. They explain the different algorithms, mention their time complexity and list notable papers that employ each algorithm to IDS. Hodo *et al.* [8] extend the ML discussion, the authors focus on the role that feature selection plays in the overall training and performance evaluation of ML techniques. An extensive discussion of features and how they impact the design and accuracy of IDS is plotted by Varma *et al.* [19]. IDS characteristics are discussed by Debar *et al.* [20], as well as, Amer and Hamilton [21].

Hamed *et al.* [22] presents an overview of IDS components, listing them as (a) pre-processing/feature extraction, (b) pattern analyzer, which involves knowledge representation and learning processes, and finally (c) decision making. They briefly discuss the benefits of each learning technique.

Additional IDS aspects are considered, for example, Amit *et al.* [23] list the various problems and challenges involved with using ML in building IDSs.

The aforementioned manuscripts are further analyzed within Section III.

Other perspectives included in recent studies focus on a single network architecture. For example, Ismail Butun *et al.* [2] discusses Wireless Sensor Networks (WSN), Zhou *et al.* [24] highlights IDS in industrial process automation while Ghaffarian and Shahriari [16] study ML and Data Mining (DM) techniques for software vulnerability.

While these surveys provide valuable information on the design and the accuracy, none provide a detailed overview of the shortcomings of available datasets nor do they provide information on tools used to carry out attacks. In this manuscript, we address these shortcomings and provide detailed complementary work to build datasets that reflect current network threats. This manuscript is complementary to prior surveys by highlighting the shortcomings of current datasets and the claims of numerous studies on their abilities to detect deprecated attacks.

III. IDS AND DATASETS SURVEY

In this section, prominent datasets are summarized, and their limitations are highlighted. Furthermore, recent IDSs are analyzed, discussing the algorithms used and the datasets the IDSs were evaluated against. Moreover, trends observed in the algorithms used by research over the past decade are discussed, highlighting a clear shift in the use of specific algorithms.

A. DATASETS

Researchers depended on benchmark datasets to evaluate their results. However, currently available datasets lack real-life characteristics of recent network traffic. This is the reason that made most of the anomaly IDSs not applicable for production environments [25]. Furthermore, IDS is unable to adapt to constant changes in networks (i.e. new nodes, changing traffic loads, changing topology, etc.). Networks are constantly changing, for this reason depending solely on old datasets doesn't help the advancement of IDS. The process of generating new datasets should consider this constant change fact. For example, proposing a standard dataset generation platform with extendable functionality, would remove the burden of generating datasets from scratch and cope with concept drift in network patterns. This recommendation and others are further discussed in Section V.

Datasets could either be real (i.e. recorded from a network set-up) or synthetic (i.e. simulated or injected traffic). Synthetic attack injection could be used to either introduce attacks to an existing dataset or balance the attack classes present in a dataset. Viegas *et al.* [25] mentioned that for a dataset to be considered, it has to cover the following properties. (a) Real network traffic (similar to production ones), (b) valid, such that it has complete scenarios. (c) Labeled, classifying each record as normal or attack, (d) variant, (e) correct, (f) can be updated easily. (g) Reproducible in order to give researchers space to compare across different datasets, and finally, (h) shareable, hence it should not contain any confidential data. Additionally, Sharafaldin *et al.* [56] mentions that (i) having an appropriate documentation for the feature and dataset collection environment is an important aspect of IDS dataset. Cordero *et al.* [57] adds (j) having high quality normal and (k) excluding any disturbance or defects as further requirements for evaluation datasets. Furthermore, for NIDS evaluation dataset, functional and non-functional requirements are elaborated in [58].

In this manuscript, we also identify two problems that impact research domains using datasets, whether they are synthetic or not. i) Sharing datasets is sometimes prohibited due to the data contained, hence, the research in the area is limited. ii) Simulating real-life scenarios and associated attacks is difficult due to the number of parameters required for the model to be viable. However, this manuscript provides a list of the most used and recent datasets.

TABLE 1 summarizes the available datasets and categorizes them based on the domain they belong to. Moreover, attacks found in each are presented. Extra remarks, including the publication year, institute and attack classes details are listed in TABLE 3. These datasets cover mobile applications, Virtual Private Networks (VPN), Tor Networks, IDS, Botnet, Network Flows and IoT. Some of the mentioned datasets are presented in [56]. The evaluation includes DEFCON [59], CAIDA [60], LBNL [61], CDX [62], Kyoto [63], Twente [64], UMASS [65] and ADFA [30].

Ring *et al.* [66] comprehensively overview of NIDS datasets covering their main features, data format, anonymity, size, availability, recording environment, balancing, etc... The authors list the datasets and their corresponding values in each of the aforementioned criteria, leaving the choice for researchers to make based on their use-case and scenario. On the contrary, Gharib *et al.* [67] propose datasets score based on the attacks' coverage, protocols' coverage, metadata availability, anonymity, heterogeneity and labeling. While the authors evaluate attacks in the datasets and present a scientific comparison, the authors fail to provide a detailed analysis of the broader impact of their analysis.

Furthermore, due to the sparsity of the details supplementing the available datasets, the task of evaluating and ranking datasets would introduce unfair results. For example, a dataset that realistically represents background and attack traffic is better than a dataset that doesn't. However, there is no standard metric to evaluate how realistic the generation is, as well as, this information is not released with the dataset.

B. IDS AND ASSOCIATED DATASETS ANALYSIS

In this section, a survey of recent ML IDS is provided, analyzing the associated datasets, and their shortcomings. IEEE Xplore and Google Scholar queries were made using "Intrusion Detection System*" OR "IDS*" filtering the dates to include manuscripts published in the last decade. The filtration was made to have a wide coverage of datasets,

TABLE 1. Attacks prominent datasets.

General Purpose Networks									Brute Force			Web								
Year	Dataset	Normal	DoS	DDoS	Probe	U2R	R2L	Infiltrating/Scanning	SSH	FTP	Heartbleed	Brute Force	XSS	Sql Injection	Webshell	DVWA	Botnet	Network and Host Events	Port Scan	Meterpreter
2018	CICIDS2018 [26]	✓	✓	✓	-	-	-	✓	✓	-	✓	✓	✓	✓	-	✓	✓	-	✓	-
2017	CICIDS2017 [27]	✓	✓	✓	-	-	-	✓	✓	✓	✓	✓	✓	✓	-	-	✓	-	✓	-
2017	CIC DoS dataset [28]	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2017 & 2013	ADFA-IDS [29], [30]	✓	-	-	-	*	+	-	✓	✓	-	-	-	-	✓	-	-	-	-	✓
2017	Unified Network Dataset [31]	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
2016	DDoSTB [32]	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2015	Booters [33]	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2015	TUIDS Coordinated Scan [34]	✓	-	-	✓	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-
	TUIDS DDoS [34]	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	TUIDS Intrusion [34]	✓	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2014	Botnet dataset [35]	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-
2012	STA2018 [36]	✓	✓	✓	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-
2011	CTU-13 [37]	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-
2010	ISCXIDS2012 [38]	✓	✓	✓	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-
2009	Waikato [39]	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2007	CAIDA DDoS [40]	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1999	NSL-KDD [41]	✓	✓	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
1999	KDD'99 [42]	✓	✓	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
1998, 1999, 2000	DARPA [43]	✓	✓	-	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-

Special Purpose Networks					
Year	Dataset	IoT	VPN	Tor	SCADA
2018	Bot-IoT [44]	✓	-	-	-
2017	Anomalies Water System [45], [46]	-	-	-	✓
2017	IoT devices captures [47]	✓	-	-	-
2016	Tor-nonTor dataset [48]	-	-	✓	-
2016	VPN-nonVPN dataset [49]	-	✓	-	-
2015	4SICS ICS [50]	-	-	-	✓

Mobile Applications					
Year	Dataset	Benign	Botnet	Adware	Malware
2016	Kharon Malware Dataset [51]	-	-	-	✓
2015 - 2017	Android Adware and General Malware Dataset [52]	✓	-	✓	✓
2010 - 2014	Android Botnet dataset [53]	-	✓	-	-
2010 - 2011	Android Malware Genome [54]	-	-	-	✓
-	AndroMalShare [55]	-	-	-	✓

*: Adding new Superuser, +: C100

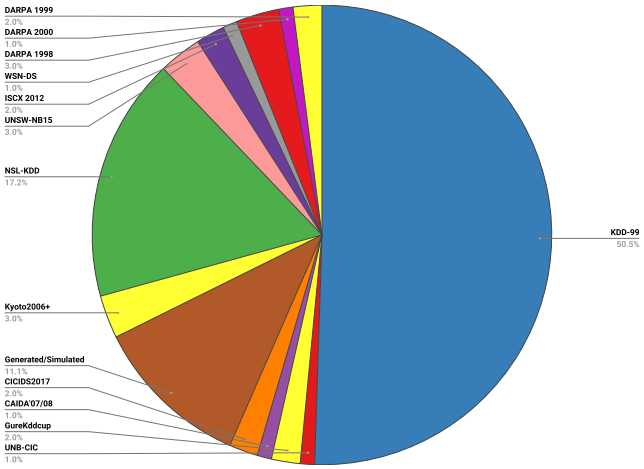


FIGURE 1. Datasets used for IDS evaluation distribution (85 IDSs manuscripts listed in Table 2).

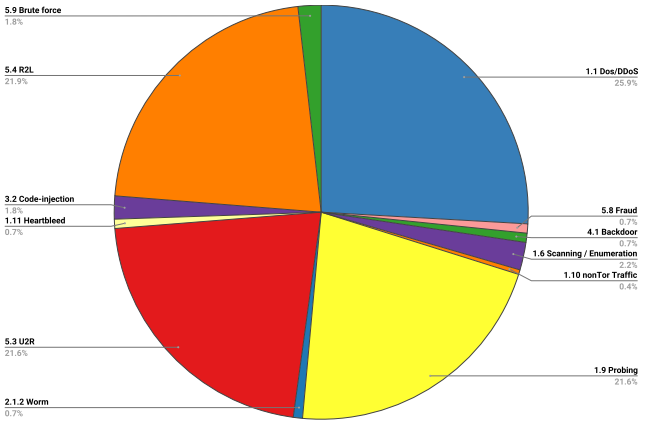


FIGURE 2. Covered attacks in discussed IDS (85 IDSs manuscripts listed in Table 2).

ML techniques and detected attacks. A total of **85** published manuscripts in the period of [2008 − 2020] are analyzed. Analysis of older IDS ML techniques and used features for the period 2004 - 2007 was previously conducted by Nguyen and Armitage [11]. They discuss the limitations of port-based and payload-based classification and the emerging use of ML techniques to classify IP traffic.

TABLE 2 summarizes the pre-eminent (i.e., most cited) IDS research from the past decade. Each IDS is mentioned with a list of the algorithms used and the datasets that the IDS was evaluated against. Moreover, the attacks detected are also listed. The algorithmic trends are then discussed alongside the attacks included in the datasets used.

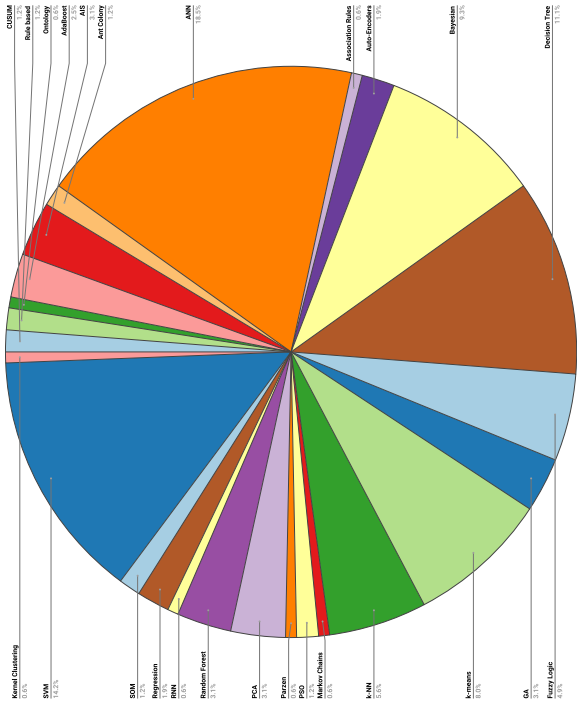
FIGURE 1 shows the distribution of datasets used for research in the last decade. Only 11% of the mentioned IDSs used generated or simulated datasets. It is also clear through this analysis that most datasets lack real-life properties, which were previously mentioned in Section III-A. FIGURE 1 also highlights the use of KDD-99 as the dataset of choice. Amjad Al Tobi and Ishbel Duncan [68] provide a comprehensive analysis of the drawbacks of the KDD'99 dataset. Moreover, Siddique *et al.* [69] provide a timeline for KDD datasets family. The provided timeline show both the different criticism points and the UCI Lab warning not to use KDD Cup'99 dataset, which further emphasizes the drawbacks of using KDD Cup'99 in the current IDS research. The second most used dataset is the DARPA datasets. DARPA datasets fail to accurately represent current attacks due to their age. Moreover, the use of the KDD'99 and DARPA datasets lead to an endemic situation, numerous results reported in literature claim detection results which are not applicable in real-world scenarios. The shortcomings of the DARPA dataset are analyzed by M. Mahoney and P. Chan [70] and John McHugh [71]. Alongside the limitations of each dataset, they are also deprecated, hence, demonstrating the inability of the IDSs presented in TABLE 2 to cope with the most recent attacks and threats.

FIGURE 2 visualizes the attacks detected by the different IDSs presented in TABLE 2. It is shown that the four attacks available in the KDD-99 dataset are the most covered, namely; DoS/DDoS, Probing, R2L, U2R. Moreover, only 12 attacks are listed in FIGURE 2 which highlights potential limitations of these IDS to cope with the broad range of attacks and zero-day attacks. To tackle the detection of zero-day attacks, there is a need to build extendable datasets that could be used to train machine learning models used for anomaly detection. By employing extendable datasets and a standardized method for dataset generation, alongside the advancement in ML [72], [73], zero-day detection could be integrated into anomaly-based IDSs. Later in Section IV, our presented threat taxonomy highlights the percentage of attack coverage achieved by current IDSs.

To further analyze the last decade research on IDSs, it is important to consider the algorithms used. Anomaly-based IDSs are based on identifying patterns that define normal and abnormal traffic. These IDSs can be classified into subcategories based on the training method used as aforementioned in Section II. These categories are identified respectively as statistical, knowledge-based and ML based. Statistical includes univariate, multivariate and time series. Knowledge-based uses finite state machines and rules like case-based, n-based, expert systems and descriptor languages. ML algorithms include artificial neural networks, clustering, genetic algorithms, Deep Learning (DL), etc.

FIGURE 3 (a) highlights the dominance of ML algorithms employed when building an IDS. As shown, both statistical and knowledge-based algorithms are less represented. This dominance is due to the significant use of ML techniques in various research domains. FIGURE 3 (a) is organized by categories (Inner Circle), subcategories (Centre Circle) and finally, the percentage of the IDSs presented in TABLE 2 using these algorithms (Outer Circle). FIGURE 3 (b) on the other hand, provides a visualization of the distribution of the algorithms used by the IDSs presented in TABLE 2. The dominance of ANN, SVM and k-means as the most used algorithms is reasoned by their ability to discriminate between

(a) Distribution of all algorithms categories



(b) Distribution of used algorithms discussed in TABLE 2

FIGURE 3. Algorithms usage distribution in the discussed IDS (85 IDSs manuscripts listed in Table 2)
Such that: AdaBoost: Adaptive Boosting
AIS: Artificial Immune System
ANN: Artificial Neural Network
CNN: Convolutional Neural Network
CUSUM: Cumulative Sum
FSM: Finite State Machine
GA: Genetic Algorithms
k-NN: k-Nearest Neighbors
ML: Machine Learning
PCA: Principal Component Analysis
PSO: Particle Swarm Optimization
RNN: Recurrent Neural Network
SOM: Self-Organizing Map
SVM: Support Vector Machine.

benign and attack classes given a feature set. However, it is important to mention that leveraging new ML techniques and adapting ones from other domains will advance the development of the next decade's IDSs.

IV. THREATS TAXONOMY

Building a generic and modular taxonomy for security threats is of high importance in order to help researchers and cyber-security practitioners build tools capable of detecting various attacks ranging from known to zero-day attacks.

Kendall *et al.* [74] proposes one of the earliest classifications of intrusions [25]. Kendall classifies intrusions into four categories namely: Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probing. In DoS, the attacker tends to prevent users from accessing a given service. When the attacker tried to gain authorized access to the target system, either by gaining local access or promoting the user to a root user, these attacks were classified as R2L and U2R respectively. Finally, probing was defined as an attacker actively footprinting a system for vulnerabilities.

Donald Welch classifies common threats in wireless networks into seven attack techniques (Traffic Analysis, Passive Eavesdropping, Active Eavesdropping, Unauthorized Access, Man-in-the-middle, Session Hijacking and Replay) [75]. In a paper by Sachin Babar *et al.* [76], the problem is addressed from a different perspective. Threats are classified according to the Internet of things security requirements (identification, communication, physical threat, embedded security and storage management). Specific domain taxonomies have also grabbed the attention of researchers. David Kotz [77] discusses privacy threats in the mobile health (mHealth) domain. In the same manner, Keshnee Padayachee [78], shows the security threats targeting compliant information and Monjur Ahmed and Alan T. Litchfield [79] works on threats from a cloud computing point of view.

This section classifies network threats based on the layers of the OSI model, provides examples of attacks for different threat types and presents a taxonomy associating network threats and the tools used to carry out attacks. The taxonomies aim at helping researchers building IDSs, but more importantly by associating the threats to the OSI model and benchmarking the threats to the tools used to carry attack or take advantage of specific vulnerabilities, the taxonomies aim to achieve higher accuracies and reduce the number of false positives of current IDS [80] and build better datasets.

A. THREAT SOURCES

FIGURE 4 identifies network threats and provides a classification according to the following criteria; (I) source of the threat, (II) affected layer based on Open Systems Interconnection (OSI) model and (III) active and passive threats. The different threats are described hereafter.

As shown, attacks can target a single layer of the OSI model, but it is important to highlight that other layers may also be affected. The taxonomy presented in this manuscript focuses on the main target layer of attack. An attack is also described to be active if it affects information, performance,

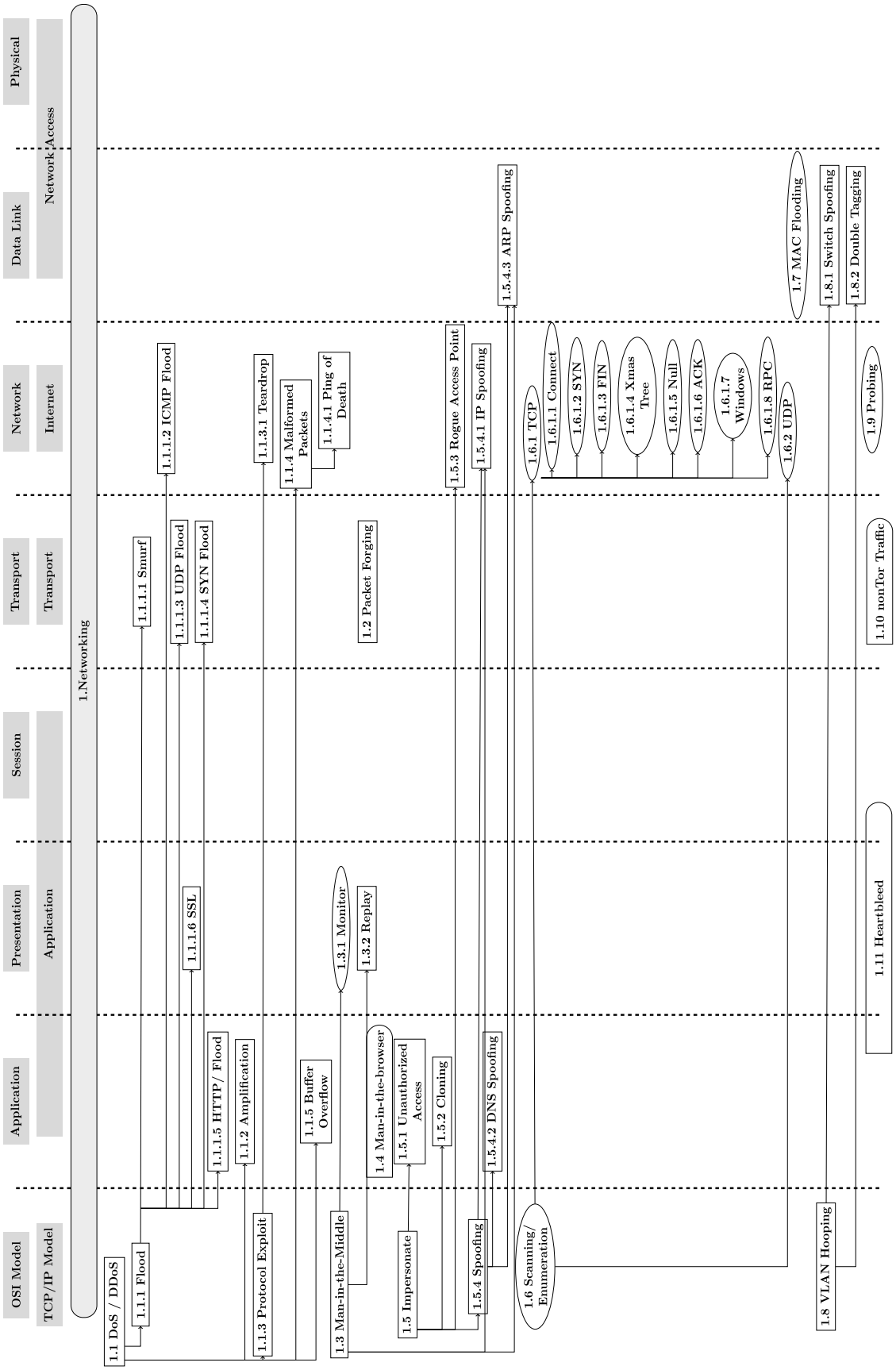


FIGURE 4. Taxonomy of threats (1 of 3).

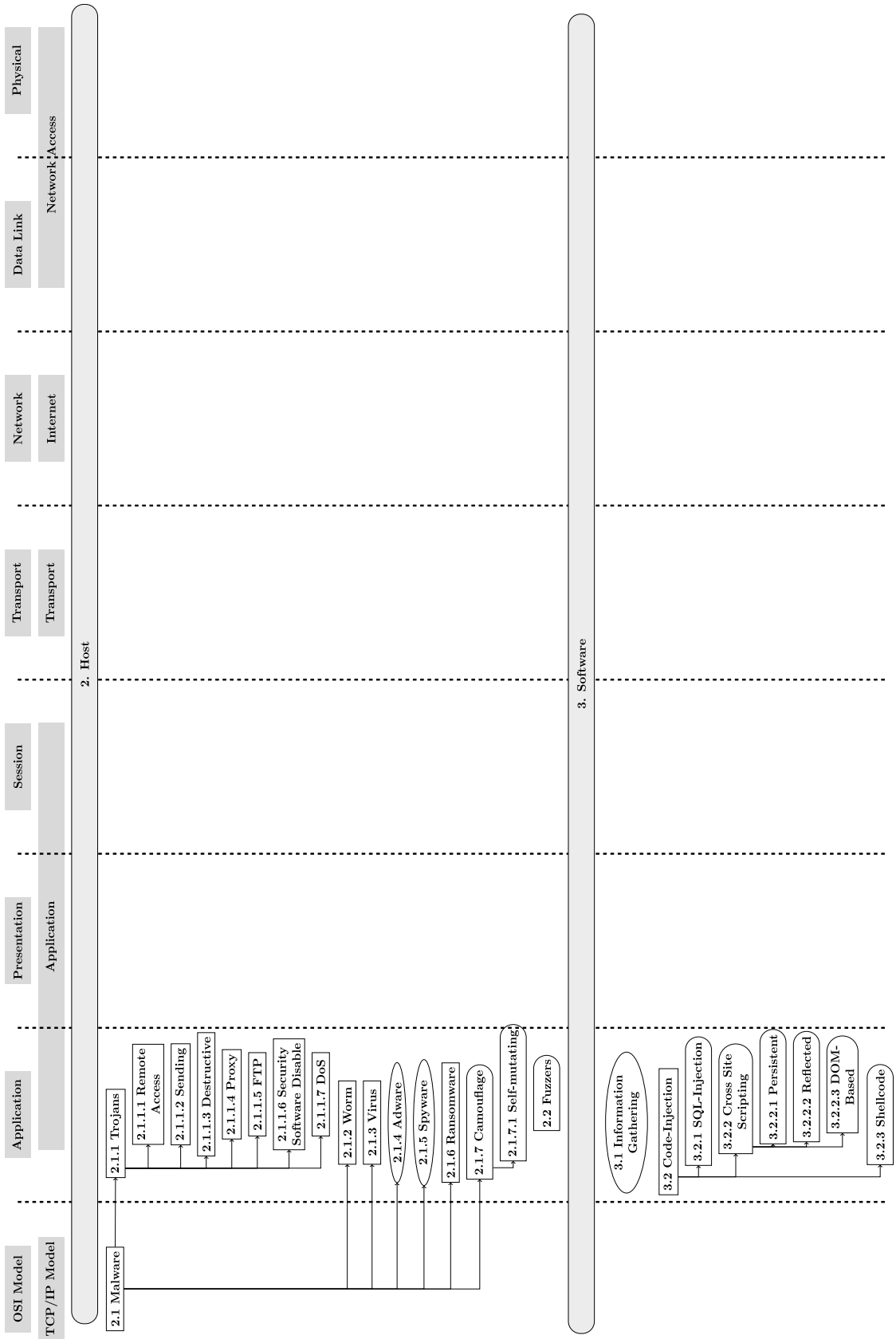


FIGURE 4. *(Continued.)* **Taxonomy of threats (2 of 3).**

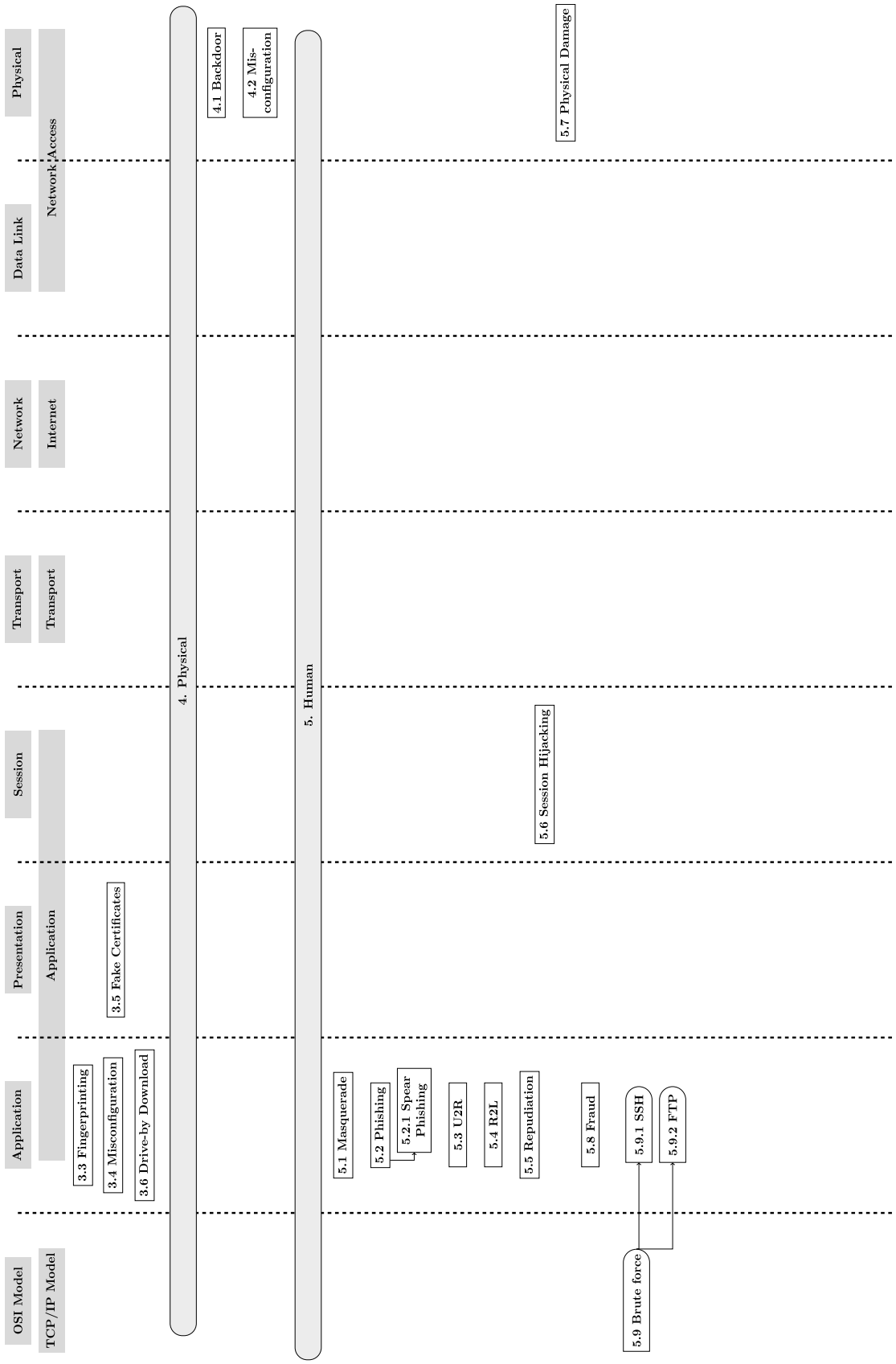


FIGURE 4. ***(Continued.)*** **Taxonomy of threats (3 of 3).**

or any aspect of the media on which it is running. In contrast to active attacks, during passive attacks the attacker is concerned with either gathering information or monitoring the network. These can be identified by their shape in FIGURE 4. Active attacks are represented by a *rectangle shape*, whilst passive attacks are represented by an *oval shape*. Attacks like adware (FIGURE 4 - 2.1.3), spyware (FIGURE 4 - 2.1.4) and information gathering (FIGURE 4 - 3.1) are considered passive attacks. DoS (FIGURE 4 - 1.1), Impersonation (FIGURE 4 - 1.4) and Virus (FIGURE 4 - 2.1.2) are forms of active attacks. However, some attacks cannot be considered active or passive until their usage is known. An example of this case is SQL-injection, if it is used for querying data from a database then it is passive. However, if it is used to alter data, drop tables or relations then the attack can be considered as active.

1) NETWORK THREATS

Threats are initiated based on a flow of packets sent over a network. Two of the most common forms of network threats are Denial of Service (DoS) and Distributed Denial of Service (DDoS) (FIGURE 4 - 1.1), where an attacker floods the network with requests rendering the service unresponsive. During these attacks, legitimate users cannot access the services. Note that common anomalies known as 'Flash Crowds' are often mistaken with DoS and DDoS attacks [81]. Flash Crowds happen when a high flow of traffic for a certain service or website occurs. This happens immediately upon the occurrence of a significant event. For example, breaking news, sales events, etc. DoS and DDoS can be divided into four categories including flood attacks (FIGURE 4 - 1.1.1), amplification attacks (FIGURE 4 - 1.1.2), protocol exploit (FIGURE 4 - 1.1.3), and malformed packets (FIGURE 4 - 1.1.4). These are defined respectively through attack examples. Smurf attacks (FIGURE 4 - 1.1.1.1) depends on generating a large amount of ping requests. Overflows (FIGURE 4 - 1.1.1.2) occurs when a program writes more bytes than allowed. This occurs when an attacker sends packets larger than 65536 bytes (allowed in the IP protocol) and the stack does not have an appropriate input sanitation in place. The ping of Death (FIGURE 4 - 1.1.4.1) attack occurs when packets are too large for routers and splitting is required. The Teardrop (FIGURE 4 - 1.1.3.1) attack takes place when an incorrect offset is set by the attacker. Finally, the SYN flood (FIGURE 4 - 1.1.1.3) attack happens when the host allocates memory for a huge number of TCP SYN packets.

Packet forging (FIGURE 4 - 1.2) is another form of networking attack. Packet forging or injection is the action where the attacker generates packets that look the same as normal network traffic. These packets can be used to perform unauthorized actions and steal sensitive data like: login credentials, personal data, credit card details, Social Security Numbers (SSN) numbers, etc. When the attacker passively monitors or intercepts communications between two or more entities and starts to control the communication, this attack is referred to as a 'Man in the Middle' attack (FIGURE 4 - 1.3). Unlike 'Man in the Middle' attack, a 'Man In The Browser' attack (1.4) intercepts the browser to alter or add fields to a web page asking the user to enter confidential data. Impersonation (FIGURE 4 - 1.5) or pretending to be another user can take different forms. The attacker may impersonate a user to gain higher security level and gain access to unauthorized data (FIGURE 4 - 1.5.1) or use cloning (FIGURE 4 - 1.5.2). Cloning is a common attack in social networks to impersonate an individual to leverage information. Rogue access points (FIGURE 4 - 1.5.3) are other impersonation forms in wireless networks. During an IP spoofing attack (FIGURE 4 - 1.5.4.1) an attacker spoofs an IP address and sends packets impersonating a legitimate host. DNS spoofing - also known as DNS cache poisoning (FIGURE 4 - 1.5.4.2) is another type of spoofing. The attacker redirects packets by poisoning the DNS. Finally, ARP spoofing (FIGURE 4 - 1.5.4.3) is used to perform attacks like Man In the Middle, in order to dissociate legitimate IP and MAC addresses in the ARP tables of victims.

Scanning/enumeration are an essential step for initiating attacks. During scanning (FIGURE 4 - 1.6), the attacker starts with searching the network for information such as: active nodes, running operating systems, software versions, etc. As defined in [82], scanning has many forms, using protocols such as TCP (FIGURE 4 - 1.6.1) or UDP (FIGURE 4 - 1.6.2). The last two examples of network attacks are media access control (MAC) address flooding (FIGURE 4 - 1.7), and VLAN hopping attack (FIGURE 4 - 1.8). In MAC flooding (FIGURE 4 - 1.7), the attacker is targeting the network switches and as a result, packets are redirected to the wrong physical ports, while the VLAN hopping attack has two forms of either switch spoofing (FIGURE 4 - 1.8.1) or double tagging (FIGURE 4 - 1.8.2).

2) HOST THREATS

Host attacks target specific hosts or systems by running malicious software to compromise or corrupt system functionalities. Most host attacks are categorized under the malware (FIGURE 4 - 2.1) category. This includes worms, viruses, adware, spyware, Trojans and ransomware. Viruses are known to affect programs and files when shared with other users on the network, whilst worms are known to self-replicate and affect multiple systems. Adware is known for showing advertisements to users when surfing the Internet or installing software. Although adware is less likely to run malicious code, it can compromise the performance of a system. Spyware gathers information such as documents, user cookies, browsing history, emails, etc. or monitors and tracks user actions. Trojans often look like trusted applications, but allow an attacker to control a device. Furthermore, camouflage malware (FIGURE 4 - 2.1.7) evolved over time reaching polymorphic and metamorphic techniques in 1990 and 1998 respectively [83], [84]. For example, self-mutating malware could use numerous techniques, such as, instruction substitution or permutation, garbage insertion, variable

substitutions and control-flow alteration [85]. Last, ransomware is a relatively new type of malware where the system is kept under the control of the attacker - or a third entity - by encrypting files until the user/organization pays a ransom [86].

3) SOFTWARE THREATS

Code injection (FIGURE 4 - 3.2) can include SQL Injection to query a database, resulting in obtaining confidential data, or deleting data by dropping columns, rows or tables. Cross-site scripting (XSS) is used to run malicious code to steal cookies or credentials. XSS has three main categories. The first is persistent/stored XSS (FIGURE 4 - 3.2.2.1), in this case, a script is saved to a database and is executed every time the page is loaded. The second is Reflected XSS (FIGURE 4 - 3.2.2.2), where the script is part of a HTTP request sent to the server. The last is DOM-based XSS (FIGURE 4 - 3.2.2.3) which can be considered as an advanced type of XSS. The attacker changes values in the Document Object Model (DOM) e.g. document location, document URL, etc. DOM-based XSS is difficult to detect as the script is never transferred to the server. Drive-by or download (FIGURE 4-3.6) is another software threat that requires no action from the user, however, the malicious code is automatically downloaded. It contributed to 48% of all web-based attacks in 2017 [87], [88] and is considered one of the main threats in 2019 [89]. Fingerprinting (FIGURE 4 - 3.3) and misconfiguration are also forms of software threats. Fake server certificates (FIGURE 4 - 3.5) are considered alarming and should be considered while analyzing communications as they could deceive the browser/user thinking that the connection is secure. This could result in phishing websites looking legitimate. Moreover, they could be used as a seed to perform other attacks like Man-in-the-Middle.

4) PHYSICAL THREATS

Physical attacks are a result of a tempering attempt on the network hardware (edge, or other devices) or its configuration. This can include changing configurations (FIGURE 4 - 4.2) and introducing backdoors (i.e. The Evil Maid).

5) HUMAN THREATS

The last category of networking attacks is one based on human actions. These include user masquerade (FIGURE 4 - 5.1). Phishing is another form of human attacks where the attacker uses emails or other electronic messaging services to obtain credentials or confidential data. When a user attempts to obtain higher privileges, it is considered a human attack like User to Root (FIGURE 4 - 5.3) and Remote to Local R2L (FIGURE 4 - 5.4). Additionally, a user can be denied an action such as repudiation attack (FIGURE 4 - 5.5). Human attacks can also include session hijacking or sniffing, these attacks are based on the attacker gaining access over an active session to access cookies and tokens.

Based on the taxonomy discussed in FIGURE 4 and the recent IDSs discussed in Section III-B, it can be seen that there are many threats that are not addressed by recent IDSs. FIGURE 5 visualizes all the threats mentioned in the taxonomy. The associated percentage represents attacks covered by the IDSs discussed in TABLE 2. As shown a large number of attacks (72%) are not covered. Hence, the network threat taxonomy aims at addressing the following:

- Help researchers generate datasets that cover non-addressed attacks.
- Provide an up-to-date taxonomy of attacks allowing to measure threats covered by datasets and the ability of IDSs to detect these threats
- Provide a structured way to address and represent threats and attacks.

B. ATTACKING TOOLS

Many tools [82], [90] have been developed to initiate different attacks. FIGURE 6 shows the main tools classified by the attacks they are used for. This can be used by researchers when building an IDS for a specific threat, then the associated tools are ones of interest. For example, for an IDS classifying impersonation attacks, Caffe-Latte, Hirte, EvilTwin and Cain and Abel are tools to check. Yaga and SQL attacks are tools used for U2R and so on.

V. CHALLENGES AND RECOMMENDATIONS

In this section, our findings are outlined based on the discussion in Section III and Section IV. A list of limitations is reviewed then the recommendations are listed.

A. LIMITATIONS AND CHALLENGES

The limitations and challenges in datasets used in IDSs can be summarized in the following:

- **Attacks Coverage**: As shown in this work only 33.3% of known attacks are covered in publicly available datasets reviewed. This is considered one of the biggest challenges preventing IDSs to be used in real-life environments.
- **Real-life Simulation**: Only 11% of the past decade IDSs use recent and/or real-life generated or simulated datasets. This demonstrates a flaw in the development of IDSs but highlights their limited ability to cope with the emerging needs.
- **Zero-Day Attacks Handling**: Attacks evolve at a pace that datasets are not currently coping with. New dataset generation techniques are needed. If the process of generating datasets and making them publicly available is made more efficient, IDS models can be quickly updated and re-trained to cope with the changes.
- **Special Purpose Datasets**: There are a limited number of available datasets serving special purpose IDSs. For example, publicly available datasets for IoT, SCADA and Tor networks are currently insufficient.
- **Dataset Outlook**: Rapid advances in networking and associated technologies require a shift in dataset

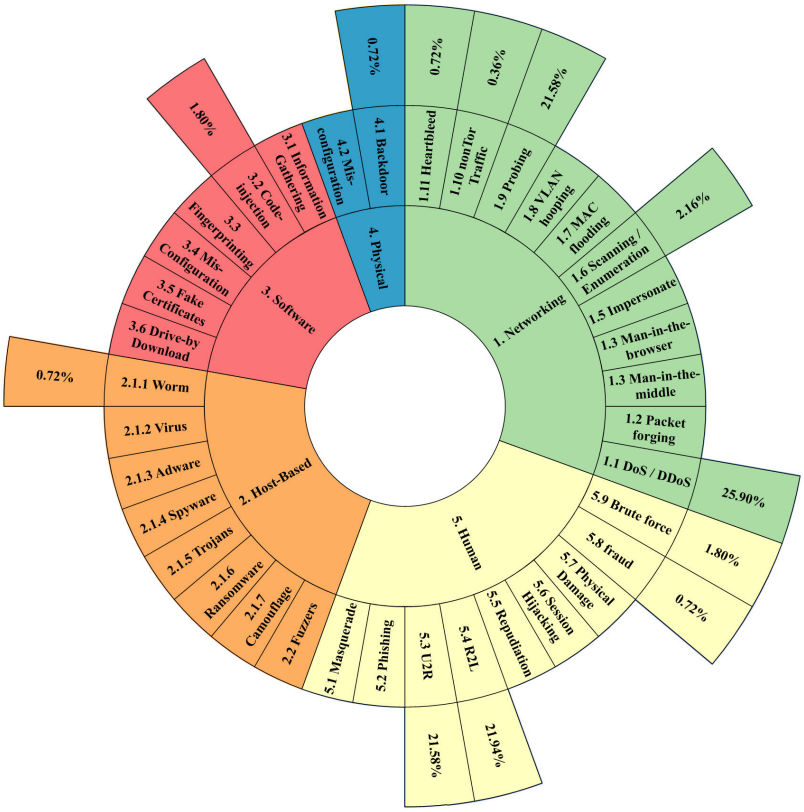


FIGURE 5. Distribution of covered attacks in discussed IDS (85 IDSs manuscripts listed in Table 2).

generation paradigm. Emerging technologies, such as Blockchain, Software Defined Network (SDN), Network Function Virtualisation (NFV), Big-Data, and their associated threats are currently not covered within available datasets. Yielding the dataset generation following trends in technologies [91]–[93].

B. RECOMMENDATIONS

Guided by the critical impact of datasets on the evolvement of IDSs and the importance of robust and accurate IDS models, the following recommendations help build the next generation IDSs. The research direction should focus equally on building complex models for IDSs and gathering/generating data that represent benign and attack scenarios accurately. This will result in IDSs suitable for real-life deployments.

- **ML-First Vs Data-First**: As discussed in Section III-A, obtaining valid, representative, and accurate data should be considered as the primary focus of research for the creation of IDSs. Building IDSs based on skewed and biased data only produces models unfit for exploitation, hence, Data-First models must be considered before ML-First.
- **Using precise evaluation metrics**: As discussed in Section II, metrics - other than accuracy - should be considered to precisely reflect an IDS performance. For example, FP and Recall should be reported. Furthermore, the geometric mean should be used with imbalanced datasets, as well as, networking metrics such as throughput. Conventional ML models report loss and accuracy by default unless other parameters are defined. Relying on the recorded loss and accuracy without measuring proposed evaluation metrics may result in misleading assessments of the overall IDS efficiency.
- **Introduce modular and extendable datasets**: As aforementioned, special purpose datasets are demanded, either to cover bespoke networks and architectures (e.g. IoT, SCADA, Tor, etc.) or to introduce new and zero-day attacks. To increase the impact of datasets, they are required to be easily extendable and capable of integrating with other datasets. As a result, datasets would be adaptable to the continuous network changes. Also, dataset generation could be rendered in the IDS pipeline, therefore, not requiring the generation of a new dataset with every introduced change. To this end, anomaly based IDSs could be trained to use advanced ML techniques to identify new and zero-day attacks.
- **Standardize attack dataset generation/collection method**: One of the main challenges forcing researchers to work with outdated datasets is the lack of documentation associated with newly available datasets. Moreover, publishing raw packet data, not only the computed features, is needed to expand the use of datasets. One of the

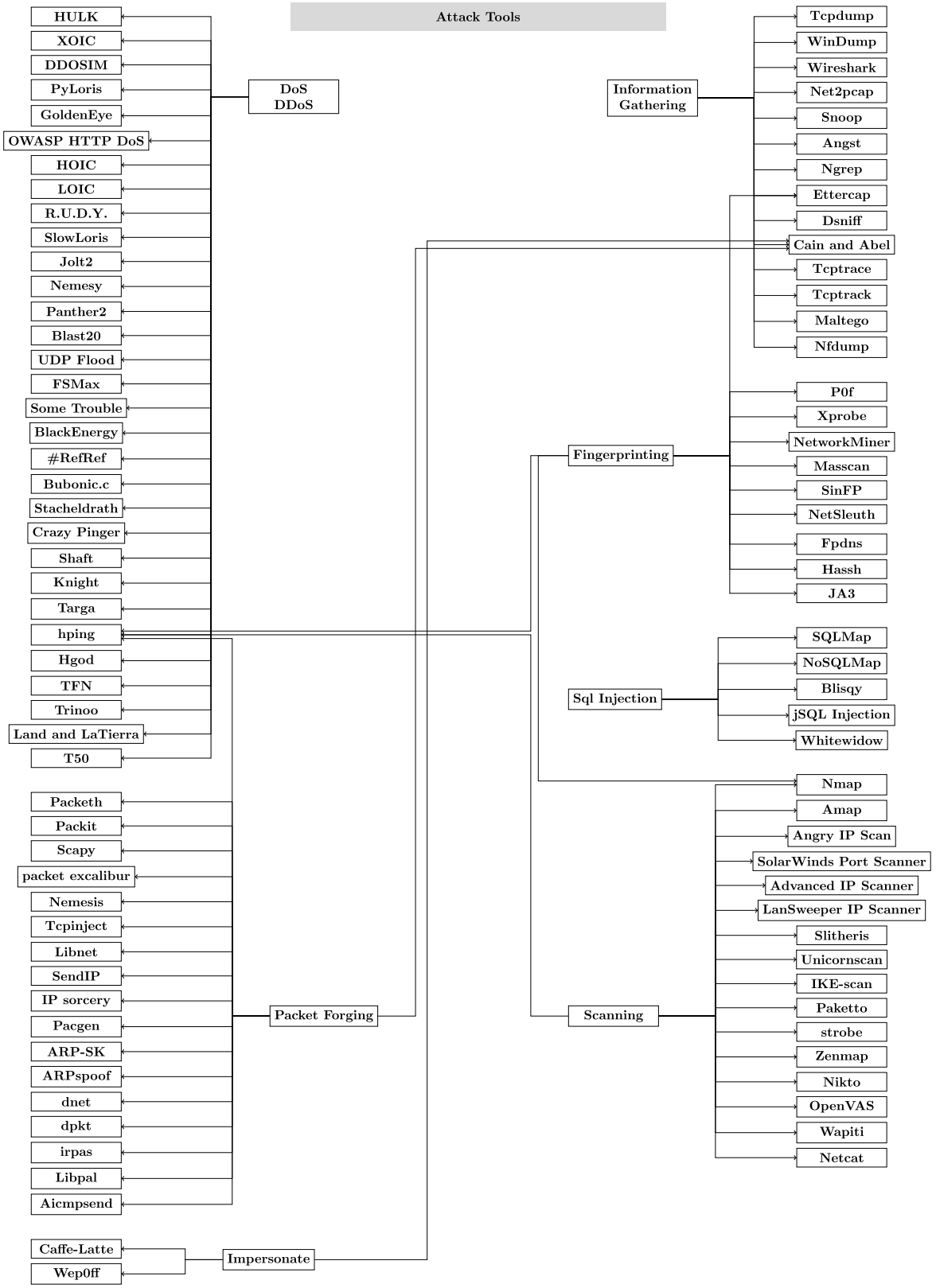


FIGURE 6. Attacking tools.

ways to generate datasets relies on α (using descriptive language to describe attacks) and β (using behavior and statistical measures to describe attacks) profiles, as described by Shiravi *et al.* [94]. Ring *et al.* [66] recommends being careful with anonymization and choosing which fields to be discarded.
Privacy is depicted as one of the main obstacles against the collection of new attacks data. Furthermore, the lack of standard tools for data collection, anonymization, documentation and publication demand researchers to use their own tailored methods.

- **Introduce models to inject realistic attacks**: Flow-Level Anomaly Modeling Engine *FLAME* [95] was one of the first tools to inject attacks that leave traces into network flows. *ID2T* [57], [96], [97] is a proposed flexible model to inject scenarios to existing datasets.
  The absence of thorough documentation of datasets makes it harder to map one dataset to another, rendering it impractical to add new attacks to existing datasets. Moreover, assuming that there exists a standard method to export realistic traffic, there isn't enough information about how to inject newly collected data to existing ones. Furthermore, the existing traffic injection proposals are limited to a single usage/prototype.
- **Generated dataset resilience**: To ensure dataset resilience, variations of the dataset should be generated. This could include variation of attack scenarios, different attacks, diverse benign traffic load (different time of day load). This, with other measures, would guarantee an extended dataset lifetime. Moreover, the complexity of dataset variation generation should be kept minimal. It is key for a dataset to be provided in a raw state to allow researchers to make a choice between using stateless or stateful analysis. If a dataset is provided only in a pre-processed form, researchers may lose the ability to use stateless or stateful detection methods and hinder the detection and accuracy of their algorithms.
- **Leveraging network monitoring to create real traffic**: Since most of the available benchmark datasets lack real-life properties, new datasets generation should benefit from network monitoring to generate realistic background traffic [34]. Furthermore, if real traffic could not be included in the dataset due to privacy concerns, real traffic should act as the ground truth for further traffic generation/simulation. Moreover, for special purpose networks, relying on released IoT/Critical infrastructure network architecture case studies should act as a guidance for network simulation. This should reduce the gap in terms of accurate realistic datasets. Different validation approaches should be used for IDS, however, with the advancement of IDS research, realistic experiments should be the main focus as demonstrated in [6].
- **Dataset Validation:** Newly generated datasets should be validated based on network traffic validation techniques. Molnár *et al.* [98] list the various metrics that could be used for this purpose. Furthermore, the similarities between real and synthetic traffic should also be evaluated as proposed in [99], [100].
- **Updated Threats Taxonomy**: The networking threat taxonomy presented at this work aims at helping create datasets that cover a wider range of attacks. Maintaining the taxonomy in a timely manner will keep it an up-to-date reference for future IDSs research. Furthermore, the taxonomy is made available for public contribution through a GitHub repository to encourage contributions from other researchers to extend, revise and update it.

While these recommendations might appear trivial at first, the majority of recent and old datasets proposed online do not conform to these guidelines as demonstrated within the previous sections. Hence, through this section, we provided recommendations for future datasets to follow ensuring the creation/generation of usable/accurate datasets.

VI. CONCLUSION

This research aims at tackling the problem of having a generic taxonomy for network threats. A proposed taxonomy is presented for categorizing network attacks based on the source, OSI model layer and whether the threat is active or passive. The prominent IDS research over the past decade (2008 - 2020) is analyzed. The analysis results in three main findings. First, benchmark datasets lack real-world properties and fail to cope with constant changes in attacks and network architectures, thus, limiting the performance of IDS. Second, we present a taxonomy of tools and associated attacks, and demonstrate that current IDS research only covers around 33.3% of threats presented in the taxonomy. Third, we highlight that - whilst ML is used by 97.25% of the examined IDS - ANN, k-means and SVM represent the majority of the algorithms used. While these algorithms present outstanding results, we also highlight that these results are obtained on outdated datasets and, therefore, not representative of real-world architectures and attack scenarios.

Finally, the network threat taxonomy and the attacks and associated tool taxonomy are open-sourced and available through GitHub,[1] allowing both security and academic researchers to contribute to the taxonomy and ensure its relevance in the future.

APPENDIX

In this appendix, Table 2 shows the prominent research over the past decade (2008 - 2020) used in the analysis presented in Section III. Each row represents one manuscript, highlighting the dataset and algorithms used within the research, alongside the attacks that the IDS is capable of detecting. Table 3 summarizes the publication year and attacks remarks for datasets discussed in Section III-A.

[1]https://github.com/AbertayMachineLearningGroup/network-threats-taxonomy

TABLE 2. Over a decade of intrusion detection systems (2008 - 2020).

Year	Dataset	Used Algorithms	Detected Attacks	Ref
2008	KDD-99	- Tree Classifiers - Bayesian Clustering	Probing, DoS, R2L, U2R	[101]
2008	KDD-99	- Parzen Classifier - v-SVC - k-means	Probing, DoS, R2L, U2R	[102]
2008	1) PIERS 2) Emergency Department Dataset 3) KDD-99	APD - Bayesian Network Likelihood - Conditional Anomaly Detection - WSARE	1) Illegal activity in imported containers 2) Anthrax 3) DoS and R2L	[103]
2008	KDD-99	- AdaBoost	Probing, DoS, R2L, U2R	[104]
2009	KDD-99	- ABC - Fuzzy Association Rules	Probing, DoS, R2L, U2R	[105]
2009	Collected transactions dataset	- Fuzzy Association Rules	Credit Card Fraud	[106]
2009	KDD-99	- Genetic-based	Probing, DoS, R2L, U2R	[107]
2009	KDD-99	- C4.5	Probing, DoS, R2L, U2R	[108]
2009	KDD-99	BSPNN using: - AdaBoost - Semi-parametric NN	Probing, DoS, R2L, U2R	[109]
2009	1999 DARPA	- RBF - Elman NN	Probing, DoS, R2L, U2R	[110]
2009	1999 DARPA	- SNORT - Non-Parametric CUSUM - EM based Clustering	13 Attack Types	[111]
2010	KDD-99	FC-ANN based on: - Fuzzy Clustering - ANN	Probing, DoS, R2L, U2R	[112]
2010	KDD-99	- Logistic Regression	Probing, DoS, R2L, U2R	[113]
2010	KDD-99	- FCM Clustering - NN	Probing, DoS, R2L, U2R	[114]
2011	Generated dataset	- OCSVM	- Nachi / Netbios scan - DDoS UDP/ TCP flood - Stealthy DDoS UDP flood - DDoS UDP flood + traffic deletion Popup spam - SSH scan + TCP flood	[115]
2011	KDD-99	- AdaBoost - NB	Probing, DoS, R2L, U2R	[116]
2011	KDD-99	- Genetic Algorithm - Weighted k-NN	DoS / DDoS	[117]
2011	KDD-99	Genetic Fuzzy Systems based on: - Michigan - Pittsburgh - IRL	Probing, DoS, R2L, U2R	[118]
2011	KDD-99	- DT - Ripper Rule - BON - RBF NN - Bayesian Network - NB	- Probing - DoS	[119]
2011	KDD-99	- K-means clustering - SOM	Probing, DoS, R2L, U2R	[120]
2011	KDD-99	- Rule-Based - BON - ART Network	Probing, DoS, R2L, U2R	[121]

TABLE 2. *(Continued.)* **Over a decade of intrusion detection systems (2008 - 2020).**

Year	Dataset	Used Algorithms	Detected Attacks	Ref
2011	KDD-99	- SVM	Probing, DoS, R2L, U2R	[122]
2011	KDD-99	- K-Means - NB	Probing, DoS, R2L, U2R	[123]
2012	KDD-99	- Modified SOM - k-means	Probing, DoS, R2L, U2R	[124]
2012	1998 DARPA	- SVM	Attack and Non-Attack	[125]
2012	1998 DARPA	ELMs: - Basic - Kernel-Based	Probing, DoS, R2L, U2R	[126]
2012	1998 DARPA	- SVDD	U2R	[127]
2012	KDD-99	- Hidden NB	Probing, DoS, R2L, U2R	[128]
2012	KDD-99	- SVM - DT - SA	Probing, DoS, R2L, U2R	[129]
2012	KDD-99	Ensemble DTs: - Decision Stump - C4.5 - NB Tree - RF - Random Tree - Representative Tree model	Probing, DoS, R2L, U2R	[130]
2012	KDD-99	- K-means - SVM - Ant Colony	Probing, DoS, R2L, U2R	[131]
2013	KDD-99	- Fuzzy C means - Fuzzy NN / Neurofuzzy - RBF SVM	Probing, DoS, R2L, U2R	[132]
2013	NSL-KDD	- Fuzzy Clustering NN	Probing, DoS, R2L, U2R	[133]
2013	KDD-99	- K-means - NN MLP	Probing, DoS, R2L, U2R	[134]
2013	KDD-99	- FFNN - ENN - GRNN - PNN - RBNN	Probing, DoS, R2L, U2R	[135]
2013	DARPA 2000	APAN using: - Markov Chain - Kmeans Clustering	DDoS	[136]
2013	ISCX 2012	KMC+NBC - K-Means Clustering - NB Classifier	Normal and Attack	[137]
2013	Bank's Credit Card Data	- DT	Fraud	[138]
2013	KDD-99	Two variants of GMDH: - Monolithic - Ensemble-based	Probing, DoS, R2L, U2R	[139]
2013	Simulated dataset	- Non-Parametric CUSUM	Jamming	[140]
2014	- KDD-99	- ELM	Probing, DoS, R2L, U2R	[141]
2014	- KDD-99 - NSL-KDD	ANN-Bayesian Net-GR ensemble: - ANN - Bayesian Net with GR feature selection	Probing, DoS, R2L, U2R	[142]
2014	NSL-KDD	- One-class SVM - C4.5 DT	-	[143]
2014	KDD-99	- K-medoids	Probing, DoS, R2L, U2R	[144]
2014	KDD-99	- SVM - CSOACN	Probing, DoS, R2L, U2R	[145]
2014	NSL-KDD	- AIS (NSA, CSA, INA)	Normal and abnormal	[146]
2015	KDD-99	- DT - CFA (Feature Selection)	Probing, DoS, R2L, U2R	[147]
2015	gureKddcup6percent	- SVM	R2L	[148]
2015	KDD-99	- K-means - k-NN	Probing, DoS, R2L, U2R	[149]

TABLE 2. *(Continued.)* **Over a decade of intrusion detection systems (2008 - 2020).**

Year	Dataset	Used Algorithms	Detected Attacks	Ref
2015	KDD-99	- Weighted ELM	Probing, DoS, R2L, U2R	[150]
2015	GureKddcup	- AIS (R-chunk)	Normal and abnormal	[151]
2016	KDD-99	- PCA - Fuzzy PCA - k-NN	Probing, DoS, R2L, U2R	[152]
2016	NSL-KDD	- PCA - SVM - MLP - C4.5 - NB	Probing, DoS, R2L, U2R	[153]
2016	Simulated dataset	- ANN	DoS/DDoS	[154]
2016	Generated dataset using httperf	- Mapping	- SQL Injection - XSS	[155]
2016	KDD-99	- SVM - PCA	- Normal and Attack	[156]
2016	NSL-KDD	- AIS (NSA-GA) - SVM - NB - DT (J48)	Normal and abnormal	[157]
2017	- Kyoto2006+ - NSL-KDD	- Forked VAE - Unsupervised deep NN	Probing, DoS, R2L, U2R	[158]
2017	KDD-99	- Binary PSO - k-NN	Probing, DoS, R2L, U2R	[159]
2017	KDD-99	- R-tree - k-NN - K-means - SVM	Probing, DoS, R2L, U2R	[160]
2017	Generated dataset	- GPU-based ANN - BON	Normal and Attack	[161]
2017	NSL-KDD	- DL RNN	Probing, DoS, R2L, U2R	[162]
2017	NSL-KDD	- K-means - NB - Information Gain	Probing, DoS, R2L, U2R	[163]
2017	UNB-CIC	- ANN - SVM	nonTor Traffic	[164]
2017	KDD-99	- Polynomial Feature Correlation	- DoS	[165]
2017	KDD-99	- PCA - k-NN - Softmax Regression	Probing, DoS, R2L, U2R	[166]
2017	KDD-99	Optimized Backpropagation by Conjugate Gradient algorithm - Fletcher Reeves - Polak Ribiere - Powell Beale	Probing, DoS, R2L, U2R	[167]
2018	KDD-99	- Kernel Clustering	Probing, DoS, R2L, U2R	[168]
2018	Simulated Dataset	- MLP - SVM - J48 - NB - Logistic - RF Features Selection: - BFS-CFS - GS-CFS	Individual and Combination Routing Attacks: - Hello Flood - Sinkhole - Wormhole	[169]
2018	KDD-99	- FLN - PSO	Probing, DoS, R2L, U2R	[170]
2018	NSL-KDD - UNSW-NB15	- Deep Auto-Encoder - ANN	Probing, DoS, R2L, U2R	[171]
2018	-KDD-99 -NSL-KDD	- DL - NDAE - Stacked NDAEs	Probing, DoS, R2L, U2R	[80]
2018	KDD-99	- KFRFS - NB - IBK - AdaBoost - MFNN - SMO - RF	Probing, DoS, R2L, U2R	[172]

TABLE 2. *(Continued.)* **Over a decade of intrusion detection systems (2008 - 2020).**

Year	Dataset	Used Algorithms	Detected Attacks	Ref
2018	NSL-KDD	- AIS (NSA, CSA)	Normal and abnormal	[173]
2018	- KDD-99 - CAIDA'07/08 - Generated traffic	- AIS	DoS	[174]
2019	NSL-KDD	- NB - RF - ANN - SVM - BayesNet - DT (Enhanced J48, J48, ADTree, DecisionStump, RandomTree, SimpleCart)	Probing, DoS, R2l, U2R	[175]
2019	KDD-99	- DT - SVM (least square) Feature Selection: - FGLCC - CFA	Probing, DoS, R2l, U2R	[176]
2019	- UNSW-NB15 - CICIDS2017	- Deep FFNN - RF - Gradient Boosting Tree	- Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shell-Code & Worms - DoS, DDoS, Web-based, Brute force, Infiltration, Heartbleed, Bot & Scan	[177]
2019	- ISCX 2012 - NSL-KDD - Kyoto2006+	- IG - PCA - SVM - IBK - MLP	- Normal and Attack - Probing, DoS, R2l, U2R	[178]
2019	- KDD-99 - NSL-KDD - UNSW-NB15 - Kyoto2006+ - WSN-DS - CICIDS2017	- Deep NN - Logistic Regression - NB - k-NN - SVM - DT - AB - RF	- Probing, DoS, R2l, U2R - 4 DoS attacks (Blackhole, Grayhole, Flooding & Scheduling) - Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shell-Code & Worms - DoS, DDoS, Web-based, Brute force, Infiltration, Heartbleed, Bot & Scan	[179]
2019	- NSL-KDD	- Genetic Algorithms - Kernel ELM - DT - k-NN - MLP - SVM	Probing, Dos, R2L, U2R	[180]
2020	Generated dataset	- AdaBoost - Decision Stump - J48 - RF - SVM - MLP - NB - BayesNet	- DDoS	[181]
2020	NSL-KDD	- Deep NN	Probing, DoS, R2l, U2R	[182]
2020	- KDD-99	- Ontology - NB - DT - RF	Probing, Dos, R2L, U2R	[183]
2020	Generated dataset	- Local Outlier Factor - Isolation Forest	Port Scanning, HTTP & SSH Brute Force & SYN Flood	[184]

Where:

* ABC: Association Based Classification
* AdaBoost : Adaptive Boosting
* AIS: Artificial Immune System
* ANN: Artificial Neural Network
* APAN: Advanced Probabilistic Approach for Network-based IDS
* APD: Anomaly Pattern Detection
* ART: Adaptive Resonance Theory
* BFS-CFS: Best First Search with Correlation Features Selection
* BON: Back-Propagation Network
* BSPNN: Boosted Subspace Probabilistic Neural Network
* IG: Information Gain
* INA: Immune Network Algorithms
* IRL: Iterative Rule Learning
* KFRFS: Kernel-based Fuzzy-Rough Feature Selection
* k-NN: k-Nearest Neighbors
* MFNN: Multi-Functional Nearest-Neighbour
* MLP: Multi-Layer Perceptron
* NB: Naïve Bayes
* NDAE: Non-Symmetric Deep Auto-Encoder
* NN: Neural Network

TABLE 2. ***(Continued.)*** **Over a decade of intrusion detection systems (2008 - 2020).**

* CFA: CuttleFish Algorithm
* CSA: Clonal Selection Algorithm
* CSOACN: Clustering based on Self-Organized Ant Colony Network
* CUSUM: CUmulative SUM
* DL: Deep Learning
* DoS: Denial of Service
* DT: Decision Tree
* ELM: Extreme Learning Machine
* ENN: Elman Neural Network
* FCM: Fuzzy C-Mean
* FFNN: Feed Forward Neural Network
* FGLCC: Feature Grouping based on Linear Correlation Coefficient
* FLN: Fast Learning Network
* GMDH: Group Method for Data Handling
* GR: Gain Ratio
* GRNN: Generalized Regression Neural Network
* GS-CFS: Greedy Stepwise with Correlation Features Selection
* IBK: Instance Based Learning
* NSA: Negative Selection Algorithm
* OCSVM: One Class Support Vector Machine
* PCA: Principal Component Analysis
* PNN: Probabilistic Neural Network
* PSO: Particle Swarm Optimization
* R2L: Remote to Local
* RBF: Radial Basis Function
* RBNN: Radial Basis Neural Network
* RF: Random Forest
* RNN: Recurrent Neural Networks
* SA: Simulated Annealing
* SOM: Self-Organizing Map
* SVDD: Support Vector Data Description
* SVM: Support Vector Machine
* U2R: User to Root
* VAE: Variational Auto-Encoder
* WSARE: What's Strange About Recent Events
* XSS: Cross Site Scripting

TABLE 3. **Datasets attacks remarks.**

Dataset Name	Institute	Attacks Remarks
General Purpose Networks		
ADFA-IDS 2017 [29], [30]	Australian Defense Force Academy	-
Booters [33]	University of Twente, SURFnet, Federal University of Rio Grande do Sul	9 DDoS attacks
Botnet dataset [35]	Canadian Institute for Cybersecurity (CIC)	7 Botnet types in training set and 16 in test set
CAIDA 2007 [40]	Center for Applied Internet Data Analysis	1 hour of DDoS attack divided into 5-minute pcap files
CIC DoS dataset [28]	CIC	8 DoS attack traces
CICIDS2017 [27]		4 DoS types, Infiltration Dropbox Download and Cool disk, 14 Port Scan types
CICIDS2018 [26]		
CTU-13 [37]	CTU University	13 captures of botnet samples
DARPA [43]	MIT Lincoln Laboratory	17 DoS, 12 U2R, 15 R2L, 10 Probing and 1 Data
DDoSTB [32]	Punjab Technical University & SBS State Technical Campus	DDoS Testbed using emulated and real nodes
ISCXIDS2012 [38]	CIC	HTTP, SMTP, SSH, IMAP, POP3, and FTP Traffic
KDD-99 [42]	University of California	Covers 24 training attack types and 14 additional types in the test data
TUIDS [34]	Tezpur University	(1) TUIDS IDS dataset. (2) TUIDS Scan dataset. (3) TUIDS DDoS dataset (22 DDoS attack types)
NSL-KDD [41]	CIC	Improvement of KDD'99 dataset
STA2018 [36]	University of St Andrews	Transformation of UNB ISCX (contains 550 features)
Unified Network Dataset [31]	Los Alamos National Laboratory	90 days of Network and Host logs
Waikato [39]	RIPE Network Coordination Center	-
Special Purpose Networks		
4SICS ICS [50]	Netresec	-
Anomalies Water System [45], [46]	French Naval Academy	15 different real situations covering cyber-attacks (DoS & Spoofing), breakdown (Sensor Failure & Wrong connection), sabotage (Blocked Measures & Floating Objects)
Bot-IoT [44]	The center of UNSW Canberra Cyber	Attacks include DoS/DDoS, OS and Service Scan, Keylogging and Data Exfiltration
IoT devices captures [47]	Aalto University	represents that data of 31 smart home IoT devices of 27 different types
Tor-nonTor dataset [48]	CIC	7 traffic categories (Browsing, Email, Chat, Audio/Video-Streaming, FTP, VoIP, P2P)
VPN-nonVPN dataset [49]		14 traffic categories (VPN-VOIP, VPN-P2P, etc.) covering Browsing, Email, Chat, Streaming, File Transfer, VoIP, TraP2P
Mobile Applications		
Android Adware and General Malware Dataset [52]	CIC	1900 application (Adware, General Malware and Benign)
Android Botnet dataset [53]		1929 Botnet samples covering 14 families (AnserverBot, Bmaster, DroidDream, Geinimi, MisoSMS, NickySpy, Not Compatible, PJapps, Pletor, RootSmart, Sandroid, TigerBot, Wroba, Zitmo
Android Malware Genome [54]	North Carolina State University	More than 1,200 malware samples
AndroMalShare [55]	Botnet Research Team and Xi'an Jiaotong University	More than 85,000 Android malware samples
Kharon Malware Dataset [51]	Kharon project	7 malware deeply examined, 10 malware (one sample each) and 2 partially examined

REFERENCES

[1] Cisco. (Sep. 2017). *Cisco Visual Networking Index: Forecast and methodology, 2016–2021*. Accessed: Feb. 15, 2018. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[2] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 266–282, 1st Quart., 2014.

[3] X. J. A. Bellekens, C. Tachtatzis, R. C. Atkinson, C. Renfrew, and T. Kirkham, "GLoP: Enabling massively parallel incident response through GPU log processing," in *Proc. 7th Int. Conf. Secur. Inf. Netw. (SIN)*. New York, NY, USA: ACM, 2014, pp. 295–301.

[4] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[5] P. Owezarski, P. Berthou, Y. Labit, and D. Gauchard, "LaasNetExp: A generic polymorphic platform for network emulationand experiments," in *Proc. 4th Int. Conf. Testbeds Res. Infrastruct. Develop. Netw. Communities (TRIDENTCOM)*, 2008, p. 10.

[6] S. Behal and K. Kumar, "Trends in validation of DDoS research," *Procedia Comput. Sci.*, vol. 85, pp. 7–15, Jan. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916305105

[7] S. Axelsson, "The base-rate fallacy and the difficulty of intrusion detection," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 3, pp. 186–205, Aug. 2000, doi: 10.1145/357830.357849.

[8] E. Hodo, X. Bellekens, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Shallow and deep networks intrusion detection system: A taxonomy and survey," pp. 1–43, 2017, *arXiv:1701.02145*. [Online]. Available: http://arxiv.org/abs/1701.02145

[9] B. Atli, "Anomaly-based intrusion detection by modeling probability distributions of flow characteristics," M.S. thesis, School Elect. Eng., Aalto Univ., Espoo, Finland, Oct. 2017. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-201710307348

[10] M. E. Aminanto, R. Choi, H. C. Tanuwidjaja, P. D. Yoo, and K. Kim, "Deep abstraction and weighted feature selection for Wi-Fi impersonation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 621–636, Mar. 2018.

[11] T. T. T. Nguyen and G. Armitage, "A survey of techniques for Internet traffic classification using machine learning," *IEEE Commun. Surveys Tuts.*, vol. 10, no. 4, pp. 56–76, 2008.

[12] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[13] S. Narkhede. (2018). *Understanding AUC—ROC Curve*. Accessed: Jul. 29, 2019. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[14] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 4th Int. Conf. Mach. Learn.*, Nashville, TN, USA, vol. 97, Jul. 1997, pp. 179–186.

[15] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "TESSERACT: Eliminating experimental bias in malware classification across space and time," pp. 1–18, 2018, *arXiv:1807.07838*. [Online]. Available: http://arxiv.org/abs/1807.07838

[16] S. M. Ghaffarian and H. R. Shahriari, "Software vulnerability analysis and discovery using machine-learning and data-mining techniques: A survey," *ACM Comput. Surv.*, vol. 50, no. 4, p. 56, Nov. 2017.

[17] Y. Chen, Y. Li, X.-Q. Cheng, and L. Guo, "Survey and taxonomy of feature selection algorithms in intrusion detection system," in *Proc. Int. Conf. Inf. Secur. Cryptol.* Berlin, Germany: Springer, 2006, pp. 153–167.

[18] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE Commun. Mag.*, vol. 57, no. 5, pp. 76–81, May 2019.

[19] P. R. K. Varma, V. V. Kumari, and S. S. Kumar, "A survey of feature selection techniques in intrusion detection system: A soft computing perspective," in *Progress in Computing, Analytics and Networking*, P. K. Pattnaik, S. S. Rautaray, H. Das, and J. Nayak, Eds. Singapore: Springer, 2018, pp. 785–793.

[20] H. Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of intrusion-detection systems," *Comput. Netw.*, vol. 31, pp. 805–822, Apr. 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128698000176

[21] S. H. Amer and J. Hamilton, "Intrusion detection systems (IDS) taxonomy-a short review," *Defense Cyber Secur.*, vol. 13, no. 2, pp. 23–30, 2010.

[22] T. Hamed, J. B. Ernst, and S. C. Kremer, "A survey and taxonomy of classifiers of intrusion detection systems," in *Computer and Network Security Essentials* (Computer and Network Security Essentials), K. Daimi, Ed. Cham, Switzerland: Springer, 2018, pp. 21–39, doi: 10.1007/978-3-319-58424-9_2.

[23] I. Amit, J. Matherly, W. Hewlett, Z. Xu, Y. Meshi, and Y. Weinberger, "Machine learning in cyber-security–problems, challenges and data sets," pp. 1–8, 2018, *arXiv:1812.07858*. [Online]. Available: http://arxiv.org/abs/1812.07858

[24] C. Zhou, S. Huang, N. Xiong, S.-H. Yang, H. Li, Y. Qin, and X. Li, "Design and analysis of multimodel-based anomaly intrusion detection systems in industrial process automation," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 10, pp. 1345–1360, Oct. 2015.

[25] E. K. Viegas, A. O. Santin, and L. S. Oliveira, "Toward a reliable anomaly-based intrusion detection in real-world environments," *Comput. Netw.*, vol. 127, pp. 200–216, Nov. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128617303225

[26] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, 2018, pp. 108–116.

[27] Canadian Institute for Cybersecurity. (2017). *Intrusion Detection Evaluation Dataset (CICIDS2017)*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/ids-2017.html

[28] (2017). *CIC DoS Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/dos-dataset.html

[29] G. Creech and J. Hu. (Mar. 2017). *ADFA IDS Dataset*. Accessed: May 13, 2019. [Online]. Available: http://www.azsecure-data.org/

[30] G. Creech and J. Hu, "Generation of a new IDS test dataset: Time to retire the KDD collection," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 4487–4492.

[31] M. J. M. Turcotte, A. D. Kent, and C. Hash, "Unified host and network data set," pp. 1–16, Aug. 2017, *arXiv:1708.07518*. [Online]. Available: https://arxiv.org/abs/1708.07518

[32] S. Behal and K. Kumar, "Measuring the impact of DDoS attacks on Web services-a realtime experimentation," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 9, p. 323, 2016.

[33] J. J. Santanna, R. van Rijswijk-Deij, R. Hofstede, A. Sperotto, M. Wierbosch, L. Z. Granville, and A. Pras, "Booters an analysis of DDoS-as-a-service attacks," in *Proc. IFIP/IEEE Int. Symp. Integr. Netw. Manage. (IM)*, May 2015, pp. 243–251.

[34] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Towards generating real-life datasets for network intrusion detection," *Int. J. Netw. Secur.*, vol. 17, no. 6, pp. 683–701, 2015.

[35] Canadian Institute for Cybersecurity. (2014). *Botnet Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/botnet.html

[36] A. A. Tobi. (Sep. 2018). *STA2018*. Accessed: Oct. 10, 2018. [Online]. Available: https://github.com/elud074/STA2018

[37] S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Comput. Secur.*, vol. 45, pp. 100–123, Sep. 2014.

[38] Canadian Institute for Cybersecurity. (2012). *Intrusion Detection Evaluation Dataset (ISCXIDS2012)*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/ids.html

[39] RNC Center. (2009). *The Waikoto Internet Trace Storage Project Dataset*. Accessed: Jun. 5, 2020. [Online]. Available: https://labs.ripe.net/datarepository/data-sets/the-waikato-internet-traffic-storage-wits-passive-datasets

[40] Center for Applied Internet Data Analysis. (2007). *The CAIDA UCSD 'DDoS Attack 2007' Dataset*. Accessed: May 5, 2020. [Online]. Available: https://www.caida.org/data/passive/ddos-20070804_dataset.xml

[41] Canadian Institute for Cybersecurity. (2009). *NSL-KDD Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/nsl.html

[42] S. Hettich and S. D. Bay, *The UCI KDD Archive*. Irvine, CA, USA: Univ. of California, Department of Information and Computer Science, 1999. Accessed: Jun. 15, 2018. [Online]. Available: http://kdd.ics.uci.edu

[43] Lincoln Laboratory. (2000). *MIT Lincoln Laboratory: DARPA Intrusion Detection Evaluation*. Accessed: Jun. 15, 2018. [Online]. Available: https://www.ll.mit.edu/ideval/data

[44] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," 2018, pp. 1–40, *arXiv:1811.00701*. [Online]. Available: http://arxiv.org/abs/1811.00701

[45] P. M. Laso, D. Brosset, and J. Puentes, "Dataset of anomalies and malicious acts in a cyber-physical subsystem," vol. 14, pp. 186–191, Oct. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352340917303402

[46] H. Hindy, D. Brosset, E. Bayne, A. Seeam, and X. Bellekens, "Improving SIEM for critical SCADA water infrastructures using machine learning," in *Computer Security*. Cham, Switzerland: Springer, 2019, pp. 3–19.

[47] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, "IoT SENTINEL: Automated device-type identification for security enforcement in IoT," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 2177–2184.

[48] Canadian Institute for Cybersecurity. (2017). *Tor-nonTor Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/tor.html

[49] Canadian Institute for Cybersecurity. (2016). *VPN-nonVPN Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/vpn.html

[50] NETRESEC. *SCADA / ICS PCAP Files From 4SICS*. Accessed: May 13, 2019. [Online]. Available: https://www.netresec.com/?page=PCAP4SICS

[51] N. Kiss, J.-F. Lalande, M. Leslous, and V. Viet Triem Tong, "Kharon dataset: Android malware under a microscope," in *Learning From Authoritative Security Experiment Results*. San Jose, CA, USA: The USENIX Association, May 2016, pp. 1–12. [Online]. Available: https://hal-univ-orleans.archives-ouvertes.fr/hal-01300752

[52] Canadian Institute for Cybersecurity. (2017). *Android Adware and General Malware Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/android-adware.html

[53] Canadian Institute for Cybersecurity. (2015). *Android Botnet Dataset*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.unb.ca/cic/datasets/android-botnet.html

[54] Y. Zhou and X. Jiang, "Dissecting Android malware: Characterization and evolution," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 95–109.

[55] Botnet Research Team, Xi'an Jiaotong University. *AndroMalShare*. Accessed: Oct. 11, 2018. [Online]. Available: http://sanddroid.xjtu.edu.cn:8080

[56] I. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, "Towards a reliable intrusion detection benchmark dataset," *Softw. Netw.*, vol. 2017, no. 1, pp. 177–200, 2018.

[57] C. G. Cordero, E. Vasilomanolakis, N. Milanov, C. Koch, D. Hausheer, and M. Muhlhauser, "ID2T: A DIY dataset creation toolkit for intrusion detection systems," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2015, pp. 739–740.

[58] C. G. Cordero, E. Vasilomanolakis, A. Wainakh, M. Mühlhäuser, and S. Nadjm-Tehrani, "On generating network traffic datasets with synthetic attacks for intrusion detection," 2019, pp. 1–31, *arXiv:1905.00304*. [Online]. Available: http://arxiv.org/abs/1905.00304

[59] The Shmoo Group. (2000). *DEFCON 8, 10 and 11*. Accessed: Jun. 15, 2018. [Online]. Available: http://cctf.shmoo.com/

[60] Center for Applied Internet Data Analysis. (2016). *CAIDA Data*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.caida.org/data/index.xml

[61] L Lawrence Berkeley National Laboratory and I International Computer Science Institute. (2005). *LBNL/ICSI Enterprise Tracing Project*. Accessed: Jun. 15, 2018. [Online]. Available: http://www.icir.org/enterprise-tracing/Overview.html

[62] B. Sangster, T. J. O. Connor, T. Cook, R. Fanelli, E. Dean, C. Morrell, and G. J. Conti, "Toward instrumenting network warfare competitions to generate labeled datasets," in *Proc. CSET*. Berkeley, CA, USA: Usenix, The Advanced Computing System Association, 2009, pp. 1–6.

[63] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation," in *Proc. 1st Workshop Building Anal. Datasets Gathering Exper. Returns Secur. (BADGERS)*. New York, NY, USA: ACM, 2011, pp. 29–36.

[64] A. Sperotto, R. Sadre, F. V. Vliet, and A. Pras, "A labeled data set for flow-based intrusion detection," in *Proc. Int. Workshop IP Oper. Manage.* Berlin, Germany: Springer, 2009, pp. 39–50.

[65] S. Prusty, B. N. Levine, and M. Liberatore, "Forensic investigation of the OneSwarm anonymous filesharing system," in *Proc. 18th ACM Conf. Comput. Commun. Secur. (CCS)*. New York, NY, USA: ACM, 2011, pp. 201–214.

[66] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," pp. 1–15, 2019, *arXiv:1903.02460*. [Online]. Available: http://arxiv.org/abs/1903.02460

[67] A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An evaluation framework for intrusion detection dataset," in *Proc. Int. Conf. Inf. Sci. Secur. (ICISS)*, Dec. 2016, pp. 1–6.

[68] A. M. A. Tobi and I. Duncan, "KDD 1999 generation faults: A review and analysis," *J. Cyber Secur. Technol.*, pp. 1–37, Mar. 2018, doi: 10.1080/23742917.2018.1518061.

[69] K. Siddique, Z. Akhtar, F. Aslam Khan, and Y. Kim, "KDD cup 99 data sets: A perspective on the role of data sets in network intrusion detection research," *Computer*, vol. 52, no. 2, pp. 41–51, Feb. 2019.

[70] M. V. Mahoney and P. K. Chan, "An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection," in *Recent Advances in Intrusion Detection*, G. Vigna, C. Kruegel, and E. Jonsson, Eds. Berlin, Germany: Springer, 2003, pp. 220–237.

[71] J. McHugh, "Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln laboratory," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 262–294, Nov. 2000, doi: 10.1145/382912.382923.

[72] P. Parrend, J. Navarro, F. Guigou, A. Deruyver, and P. Collet, "Foundations and applications of artificial intelligence for zero-day and multi-step attack detection," *EURASIP J. Inf. Secur.*, vol. 2018, no. 1, p. 4, Apr. 2018, doi: 10.1186/s13635-018-0074-y.

[73] J.-Y. Kim, S.-J. Bu, and S.-B. Cho, "Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders," *Inf. Sci.*, vols. 460–461, pp. 83–102, Sep. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025518303475

[74] K. Kendall, "A database of computer attacks for the evaluation of intrusion detection systems," M.S. thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, 1999.

[75] D. Welch and S. Lathrop, "Wireless security threat taxonomy," in *Proc. IEEE Syst., Man Cybern. Society Inf. Assurance Workshop*, Jun. 2003, pp. 76–83.

[76] S. Babar, P. Mahalle, A. Stango, N. Prasad, and R. Prasad, "Proposed security model and threat taxonomy for the Internet of Things (IoT)," in *Proc. Int. Conf. Netw. Secur. Appl.* Berlin, Germany: Springer, 2010, pp. 420–429.

[77] D. Kotz, "A threat taxonomy for mHealth privacy," in *Proc. 3rd Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2011, pp. 1–6.

[78] K. Padayachee, "Taxonomy of compliant information security behavior," *Comput. Secur.*, vol. 31, no. 5, pp. 673–680, Jul. 2012.

[79] M. Ahmed and A. T. Litchfield, "Taxonomy for identification of security issues in cloud computing environments," *J. Comput. Inf. Syst.*, vol. 58, no. 1, pp. 79–88, Jan. 2018.

[80] N. Shone, T. Nguyen Ngoc, V. Dinh Phai, and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 1, pp. 41–50, Feb. 2018.

[81] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and Web sites," in *Proc. 11th Int. Conf. World Wide Web (WWW)*. New York, NY, USA: ACM, 2002, pp. 293–304.

[82] S. McClure, J. Scambray, and G. Kurtz, *Hacking Exposed 7: Network Security Secrets and Solutions*, 6th ed. New York, NY, USA: McGraw-Hill, 2009.

[83] B. B. Rad, M. Masrom, and S. Ibrahim, "Camouflage in malware: From encryption to metamorphism," *Int. J. Comput. Sci. Netw. Secur.*, vol. 12, no. 8, pp. 74–83, 2012.

[84] H. S. Galal, Y. B. Mahdy, and M. A. Atiea, "Behavior-based features model for malware detection," *J. Comput. Virol. Hacking Techn.*, vol. 12, no. 2, pp. 59–67, May 2016, doi: 10.1007/s11416-015-0244-0.

[85] D. Bruschi, L. Martignoni, and M. Monga, "Code normalization for self-mutating malware," *IEEE Secur. Privacy Mag.*, vol. 5, no. 2, pp. 46–54, Mar. 2007.

[86] Comodo. (Feb. 2018). *Malware Vs Viruses: What's the Difference?* Accessed: Feb. 28, 2018. [Online]. Available: https://antivirus.comodo.com/blog/computer-safety/malware-vs-viruses-whats-difference/

[87] A. Javed, P. Burnap, and O. Rana, "Prediction of drive-by download attacks on Twitter," *Inf. Process. Manage.*, vol. 56, pp. 1133–1145, May 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457317305824

[88] L. Neely. (2017). *2017 Threat Landscape Survey: Users on the Front Line*. SANS Institute. [Online]. Available: https://www.sans.org/reading-room/whitepapers/threats/2017-threat-landscape-surveyusers-front-line-37910

[89] SecurityFirst. (Aug. 2018). *The Top 9 Network Security Threats of 2019*. Accessed: May 16, 2019. [Online]. Available: https://securityfirstcorp.com/the-top-9-network-security-threats-of-2019/

[90] N. Hoque, M. H. Bhuyan, R. C. Baishya, D. K. Bhattacharyya, and J. K. Kalita, "Network attacks: Taxonomy, tools and systems," *J. Netw. Comput. Appl.*, vol. 40, pp. 307–324, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1084804513001756

[91] W. Meng, E. W. Tischhauser, Q. Wang, Y. Wang, and J. Han, "When intrusion detection meets blockchain technology: A review," *IEEE Access*, vol. 6, pp. 10179–10188, 2018.

[92] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based IDS and SDN," in *Proc. ACM Int. Workshop Secur. Softw. Defined Netw. Netw. Function Virtualization (SDN-NFVSec)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 13–16, doi: 10.1145/3309194.3309199.

[93] A. Derhab, M. Guerroumi, A. Gumaei, L. Maglaras, M. A. Ferrag, M. Mukherjee, and F. A. Khan, "Blockchain and random subspace learning-based IDS for SDN-enabled industrial IoT security," *Sensors*, vol. 19, no. 14, p. 3119, Jul. 2019.

[94] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, May 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404811001672

[95] D. Brauckhoff, A. Wagner, and M. May, "FLAME: A flow-level anomaly modeling engine," in *Proc. CSET*, Jul. 2008, pp. 1–6.

[96] Telecooperation Lab—TU Darmstadt. *Official ID2T Repository. ID2T Creates Labeled IT Network Datasets That Contain User Defined Synthetic Attacks*. Accessed: Sep. 5, 2019. [Online]. Available: https://github.com/tklab-tud/ID2T

[97] E. Vasilomanolakis, C. G. Cordero, N. Milanov, and M. Muhlhauser, "Towards the creation of synthetic, yet realistic, intrusion detection datasets," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2016, pp. 1209–1214.

[98] S. Molnar, P. Megyesi, and G. Szabo, "How to validate traffic generators?" in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Jun. 2013, pp. 1340–1344.

[99] F. Hernández-Campos, F. Donelson, and S. K. Jeffay, "How real can synthetic network traffic be?" in *Proc. ACM SIGCOMM (Poster Session)*. New York, NY, USA: Citeseer, 2004, p. 1.

[100] J. Cao, W. S. Cleveland, Y. Gao, K. Jeffay, F. D. Smith, and M. Weigle, "Stochastic models for generating synthetic HTTP source traffic," in *Proc. IEEE INFOCOM*, vol. 3, Mar. 2004, pp. 1546–1557.

[101] C. Xiang, P. C. Yong, and L. S. Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 918–924, May 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508000251

[102] G. Giacinto, R. Perdisci, M. Del Rio, and F. Roli, "Intrusion detection in computer networks by a modular ensemble of one-class classifiers," *Inf. Fusion*, vol. 9, no. 1, pp. 69–82, Jan. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253506000765

[103] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: ACM, 2008, pp. 169–176.

[104] W. Hu, W. Hu, and S. Maybank, "AdaBoost-based algorithm for network intrusion detection," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 38, no. 2, pp. 577–583, Apr. 2008.

[105] A. Tajbakhsh, M. Rahmati, and A. Mirzaei, "Intrusion detection using fuzzy association rules," *Appl. Soft Comput.*, vol. 9, no. 2, pp. 462–469, Mar. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494608000975

[106] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "Association rules applied to credit card fraud detection," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3630–3640, Mar. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408001176

[107] K. Shafi and H. A. Abbass, "An adaptive genetic-based signature learning system for intrusion detection," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12036–12043, Dec. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417409002589

[108] S.-Y. Wu and E. Yen, "Data mining-based intrusion detectors," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5605–5612, Apr. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408004089

[109] T. Phuoc Tran, L. Cao, D. Tran, and C. Duc Nguyen, "Novel intrusion detection using probabilistic neural network and adaptive boosting," vol. 6, no. 1, pp. 83–91, 2009, *arXiv:0911.0485*. [Online]. Available: http://arxiv.org/abs/0911.0485

[110] X. Tong, Z. Wang, and H. Yu, "A research using hybrid RBF/Elman neural networks for intrusion detection system secure model," *Comput. Phys. Commun.*, vol. 180, no. 10, pp. 1795–1801, Oct. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010465509001519

[111] W. Lu and H. Tong, "Detecting network anomalies using CUSUM and EM clustering," in *Proc. Int. Symp. Intell. Comput. Appl.* Berlin, Germany: Springer, 2009, pp. 297–308.

[112] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using artificial neural networks and fuzzy clustering," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6225–6232, Sep. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410001417

[113] M. S. Mok, S. Y. Sohn, and Y. H. Ju, "Random effects logistic regression model for anomaly detection," *Expert Syst. Appl.*, vol. 37, no. 10, pp. 7162–7166, Oct. 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410002885

[114] M. M. T. Jawhar and M. Mehrotra, "Design network intrusion detection system using hybrid fuzzy-neural network," *Int. J. Comput. Sci. Secur.*, vol. 4, no. 3, pp. 285–294, 2010.

[115] C. Wagner, J. François, and T. Engel, "Machine learning approach for IP-flow record anomaly detection," in *Proc. Int. Conf. Res. Netw.* Berlin, Germany: Springer, 2011, pp. 28–39.

[116] C. M. Rahman, D. M. Farid, and M. Z. Rahman, "Adaptive intrusion detection based on boosting and naive Bayesian classifier," *Int. J. Comput. Appl.*, vol. 24, no. 3, pp. 11–19, 2011.

[117] M.-Y. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3492–3498, Apr. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410009450

[118] M. S. Abadeh, H. Mohamadi, and J. Habibi, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7067–7075, Jun. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410013692

[119] P. Sangkatsanee, N. Wattanapongsakorn, and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches," *Comput. Commun.*, vol. 34, no. 18, pp. 2227–2235, Dec. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S014036641100209X

[120] S. Lee, G. Kim, and S. Kim, "Self-adaptive and dynamic clustering for online anomaly detection," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14891–14898, Nov. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411008426

[121] S.-S. Wang, K.-Q. Yan, S.-C. Wang, and C.-W. Liu, "An integrated intrusion detection system for cluster-based wireless sensor networks," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 15234–15243, Nov. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411008608

[122] Y. Yi, J. Wu, and W. Xu, "Incremental SVM based on reserved set for network intrusion detection," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7698–7707, Jun. 2011.

[123] Z. Muda, W. Yassin, M. N. Sulaiman, and N. I. Udzir, "A K-Means and naive bayes learning approach for better intrusion detection," *Inf. Technol. J.*, vol. 10, no. 3, pp. 648–655, Mar. 2011.

[124] A. S. Aneetha, "The combined approach for anomaly detection using neural networks and clustering techniques," *Comput. Sci. Eng., Int. J.*, vol. 2, no. 4, pp. 37–46, Aug. 2012.

[125] C. A. Catania, F. Bromberg, and C. G. Garino, "An autonomous labeling approach to support vector machines algorithms for network traffic anomaly detection," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1822–1829, Feb. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411011808

[126] C. Cheng, W. Peng Tay, and G.-B. Huang, "Extreme learning machines for intrusion detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.

[127] I. Kang, M. K. Jeong, and D. Kong, "A differentiated one-class classification method with applications to intrusion detection," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 3899–3905, Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411009286

[128] L. Koc, T. A. Mazzuchi, and S. Sarkani, "A network intrusion detection system based on a hidden Naïve Bayes multiclass classifier," *Expert Syst. Appl.*, vol. 39, no. 18, pp. 13492–13500, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412008640

[129] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Appl. Soft Comput.*, vol. 12, no. 10, pp. 3285–3290, Oct. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494612002402

[130] S. S. Sivatha Sindhu, S. Geetha, and A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 129–141, Jan. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411009080

[131] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 424–430, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417411009948

[132] A. M. Chandrashekhar and Raghuveer, "Fortification of hybrid intrusion detection system using variants of neural networks and support vector machines," *Int. J. Netw. Secur. Appl.*, vol. 5, no. 1, pp. 71–90, Jan. 2013.

[133] D. A. A. Zainaddin and Z. H. Hanapi, "Hybrid of fuzzy clustering neural network over NSL dataset for intrusion detection system," *J. Comput. Sci.*, vol. 9, no. 3, pp. 391–403, Mar. 2013.

[134] M. M. Lisehroodi, Z. Muda, and W. Yassin, "A hybrid framework based on neural network MLP and k-means clustering for intrusion detection system," in *Proc. 4th Int. Conf. Comput. Inf. (ICOCI)*, 2013, pp. 305–311.

[135] S. Devaraju and S. Ramakrishnan, "Detection of accuracy for intrusion detection system using neural network classifier," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 1, pp. 338–345, 2013.

[136] S. Shin, S. Lee, H. Kim, and S. Kim, "Advanced probabilistic approach for network intrusion forecasting and detection," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 315–322, Jan. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412009128

[137] W. Yassin, N. I. Udzir, Z. Muda, and M. N. Sulaiman, "Anomaly-based intrusion detection through k-means clustering and naives Bayes classification," in *Proc. 4th Int. Conf. Comput. Inf. (ICOCI)*, 2013, pp. 298–303.

[138] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5916–5923, Nov. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413003072

[139] Z. A. Baig, S. M. Sait, and A. Shaheen, "GMDH-based networks for intelligent intrusion detection," *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1731–1740, Aug. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095219761300050X

[140] Z. M. Fadlullah, H. Nishiyama, N. Kato, and M. M. Fouda, "Intrusion detection system (IDS) for combating attacks against cognitive radio networks," *IEEE Netw.*, vol. 27, no. 3, pp. 51–56, May 2013.

[141] J. Xiang, M. Westerlund, D. Sovilj, and G. Pulkkis, "Using extreme learning machine for intrusion detection in a big data environment," in *Proc. Workshop Artif. Intell. Secur. Workshop (AISec)*. New York, NY, USA: ACM, 2014, pp. 73–82.

[142] A. KumarShrivas and A. K. Dewangan, "An ensemble model for classification of attacks with feature selection based on KDD99 and NSL-KDD data set," *Int. J. Comput. Appl.*, vol. 99, no. 15, pp. 8–13, Aug. 2014.

[143] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1690–1700, Mar. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413006878

[144] R. Ranjan and G. Sahoo, "A new clustering approach for anomaly intrusion detection," vol. 4, no. 2, pp. 29–38, 2014, *arXiv:1404.2772*. [Online]. Available: http://arxiv.org/abs/1404.2772

[145] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks," *Future Gener. Comput. Syst.*, vol. 37, pp. 127–140, Jul. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X13001416

[146] N. Afzali Seresht and R. Azmi, "MAIS-IDS: A distributed intrusion detection system using multi-agent AIS approach," *Eng. Appl. Artif. Intell.*, vol. 35, pp. 286–298, Oct. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197614001444

[147] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2670–2679, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417414006952

[148] B. W. Masduki, K. Ramli, F. A. Saputra, and D. Sugiarto, "Study on implementation of machine learning methods combination for improving attacks detection accuracy on intrusion detection system (IDS)," in *Proc. Int. Conf. Qual. Res. (QiR)*, Aug. 2015, pp. 56–64.

[149] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowl.-Based Syst.*, vol. 78, pp. 13–21, Apr. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705115000167

[150] W. Srimuang and S. Intarasothonchun, "Classification model of network intrusion using weighted extreme learning machine," in *Proc. 12th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2015, pp. 190–194.

[151] E. A. E. R. Abas, H. Abdelkader, and A. Keshk, "Artificial immune system based intrusion detection," in *Proc. IEEE 7th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Dec. 2015, pp. 542–546.

[152] A. Hadri, K. Chougdali, and R. Touahni, "Intrusion detection system using PCA and fuzzy PCA techniques," in *Proc. Int. Conf. Adv. Commun. Syst. Inf. Secur. (ACOSIS)*, Oct. 2016, pp. 1–7.

[153] B. Subba, S. Biswas, and S. Karmakar, "Enhancing performance of anomaly based intrusion detection systems through dimensionality reduction using principal component analysis," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Nov. 2016, pp. 1–6.

[154] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson, "Threat analysis of IoT networks using artificial neural network intrusion detection system," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, May 2016, pp. 1–6.

[155] P. A. Sonewar and S. D. Thosar, "Detection of SQL injection and XSS attacks in three tier Web applications," in *Proc. Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2016, pp. 1–4.

[156] P. Nskh, M. N. Varma, and R. R. Naik, "Principle component analysis based intrusion detection system using support vector machine," in *Proc. IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2016, pp. 1344–1350.

[157] O. Igbe, I. Darwish, and T. Saadawi, "Distributed network intrusion detection systems: An artificial immune system approach," in *Proc. IEEE 1st Int. Conf. Connected Health, Appl., Syst. Eng. Technol. (CHASE)*, Jun. 2016, pp. 101–106.

[158] G. Osada, K. Omote, and T. Nishide, "Network intrusion detection based on semi-supervised variational auto-encoder," in *Computer Security—ESORICS*, S. N. Foley, D. Gollmann, and E. Snekkenes, Eds. Cham, Switzerland: Springer, 2017, pp. 344–361.

[159] A. R. Syarif and W. Gata, "Intrusion detection system using hybrid binary PSO and K-nearest neighborhood algorithm," in *Proc. 11th Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, Oct. 2017, pp. 181–186.

[160] B. Xu, S. Chen, H. Zhang, and T. Wu, "Incremental k-NN SVM method in intrusion detection," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 712–717.

[161] C. Tran, T. N. Vo, and T. N. Thinh, "HA-IDS: A heterogeneous anomaly-based intrusion detection system," in *Proc. 4th NAFOSTED Conf. Inf. Comput. Sci.*, Nov. 2017, pp. 156–161.

[162] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[163] D. A. Effendy, K. Kusrini, and S. Sudarmawan, "Classification of intrusion detection system (IDS) based on computer network," in *Proc. 2nd Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Nov. 2017, pp. 90–94.

[164] E. Hodo, X. Bellekens, E. Iorkyase, A. Hamilton, C. Tachtatzis, and R. Atkinson, "Machine learning approach for detection of nonTor traffic," in *Proc. 12th Int. Conf. Availability, Rel. Secur.* New York, NY, USA: ACM, Aug. 2017, Art. no. 85, doi: 10.1145/3098954.3106068.

[165] Q. Li, Z. Tan, A. Jamdagni, P. Nanda, X. He, and W. Han, "An intrusion detection system based on polynomial feature correlation analysis," in *Proc. IEEE Trustcom/BigDataSE/ICESS*, Aug. 2017, pp. 978–983.
[166] S. Zhao, W. Li, T. Zia, and A. Y. Zomaya, "A dimension reduction model and classifier for anomaly-based intrusion detection in Internet of Things," in *Proc. IEEE 15th Int. Conf. Dependable, Autonomic Secure Comput., 15th Intl Conf Pervas. Intell. Comput., 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congress(DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 836–843.
[167] U. N. Wisesty and Adiwijaya, "Comparative study of conjugate gradient to optimize learning process of neural network for intrusion detection system (IDS)," in *Proc. 3rd Int. Conf. Sci. Inf. Technol. (ICSITech)*, Oct. 2017, pp. 459–464.
[168] D. He, X. Chen, D. Zou, L. Pei, and L. Jiang, "An improved kernel clustering algorithm used in computer network intrusion detection," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2018, pp. 1–5.
[169] M. N. Napiah, M. Y. I. Bin Idris, R. Ramli, and I. Ahmedy, "Compression header analyzer intrusion detection system (CHA–IDS) for 6LoWPAN communication protocol," *IEEE Access*, vol. 6, pp. 16623–16638, 2018.
[170] M. H. Ali, B. A. D. Al Mohammed, A. Ismail, and M. F. Zolkipli, "A new intrusion detection system based on fast learning network and particle swarm optimization," *IEEE Access*, vol. 6, pp. 20255–20261, 2018.
[171] M. Al-Hawawreh, N. Moustafa, and E. Sitnikova, "Identification of malicious activities in industrial Internet of Things based on deep learning models," *J. Inf. Secur. Appl.*, vol. 41, pp. 1–11, Aug. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214212617306002
[172] Q. Zhang, Y. Qu, and A. Deng, "Network intrusion detection using kernel-based fuzzy-rough feature selection," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2018, pp. 1–6.
[173] D. Hooks, X. Yuan, K. Roy, A. Esterline, and J. Hernandez, "Applying artificial immune system for intrusion detection," in *Proc. IEEE 4th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2018, pp. 287–292.
[174] J. M. Vidal, A. L. S. Orozco, and L. J. G. Villalba, "Adaptive artificial immune networks for mitigating DoS flooding attacks," *Swarm Evol. Comput.*, vol. 38, pp. 94–108, Feb. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2210650216304679
[175] S. Aljawarneh, M. B. Yassein, and M. Aljundi, "An enhanced J48 classification algorithm for the anomaly intrusion detection systems," *Cluster Comput.*, vol. 22, no. S5, pp. 10549–10565, Sep. 2019, doi: 10.1007/s10586-017-1109-8.
[176] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-Ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *J. Inf. Secur. Appl.*, vol. 44, pp. 80–88, Feb. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2214212618304617
[177] O. Faker and E. Dogdu, "Intrusion detection using big data and deep learning techniques," in *Proc. ACM Southeast Conf. (ACM SE)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 86–93, doi: 10.1145/3299815.3314439.
[178] F. Salo, A. B. Nassif, and A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," *Comput. Netw.*, vol. 148, pp. 164–175, Jan. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128618303037
[179] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019.
[180] J. Ghasemi, J. Esmaily, and R. Moradinezhad, "Intrusion detection system using an optimized kernel extreme learning machine and efficient features," *Sādhanā*, vol. 45, no. 1, p. 2, Dec. 2019, doi: 10.1007/s12046-019-1230-x.
[181] S. Sen, K. D. Gupta, and M. M. Ahsan, "Leveraging machine learning approach to setup software-defined network(SDN) controller rules during DDoS attack," in *Proc. Int. Joint Conf. Comput. Intell.*, M. S. Uddin and J. C. Bansal, Eds. Singapore: Springer, 2020, pp. 49–60.
[182] Z. Liu, M.-U.-D. Ghulam, Y. Zhu, X. Yan, L. Wang, Z. Jiang, and J. Luo, "Deep learning approach for IDS," in *Proc. 4th Int. Congr. Inf. Commun. Technol.*, X.-S. Yang, S. Sherratt, N. Dey, and A. Joshi, Eds. Singapore: Springer, 2020, pp. 471–479.
[183] M. Sarnovsky and J. Paralic, "Hierarchical intrusion detection using machine learning and knowledge model," *Symmetry*, vol. 12, no. 2, p. 203, Feb. 2020.
[184] M. Eskandari, Z. H. Janjua, M. Vecchio, and F. Antonelli, "Passban IDS: An intelligent anomaly based intrusion detection system for IoT edge devices," *IEEE Internet Things J.*, early access, Jan. 30, 2020, doi: 10.1109/JIOT.2020.2970501.

HANAN HINDY (Member, IEEE) received the bachelor's degree (Hons.) and the master's degree in computer science from the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt, in 2012 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Division of Cyber-Security, Abertay University, Dundee, U.K. She is also working on utilising deep learning for IDS. Her research interests include machine learning and cyber security.

DAVID BROSSET received the master's degree in computer science from the University of South Brittany, in 2003, and the Ph.D. degree in computer science from the Arts et Metiers, Paris, in 2008. Since 2011, he has been an Associate Professor of Computer Science with the French Naval Academy. He is involved in the Chair of cyber defence of naval systems. His research interests include the domain of cyber security and in particular the cyber defence of critical systems on board.

ETHAN BAYNE received the B.Sc. degree in computing and networks, the M.Sc. degree in ethical hacking and computer security, and the Ph.D. degree in digital forensics from Abertay University, Dundee, U.K., in 2008, 2013, and 2016, respectively. He is currently a Lecturer in cyber security and computer science with the Department of Cyber Security, Abertay University. His current research interests include digital forensics, massively-parallel computation, pattern matching, machine learning, and the Internet of Things (IoT).

AMAR SEEAM (Member, IEEE) received the B.Eng. degree (Hons.) in mechanical engineering and the M.Sc. degree in information technology from the University of Glasgow, in 2003 and 2004, respectively, and the M.Sc. degree in system level integration and the Ph.D. degree from the University of Edinburgh, in 2005 and 2015, respectively. He is currently a Senior Lecturer of computer science with Middlesex University, Mauritius. His research interests include simulation assisted control, cybersecurity, the Internet of Things, and building information modeling.

CHRISTOS TACHTATZIS (Senior Member, IEEE) received the B.Eng. degree (Hons.) in communication systems engineering from the University of Portsmouth, in 2001, and the M.Sc. degree in communications, control and digital signal processing and the Ph.D. degree in electronic and electrical engineering from Strathclyde University, in 2002 and 2008, respectively. He is currently a Senior Lecturer Chancellors Fellow of sensor systems and asset management with the University of Strathclyde. He has 12 years' experience, in Sensor Systems ranging from electronic devices, networking, and communications and signal processing. His current research interests include extracting actionable information from data using machine learning and artificial intelligence.

ROBERT ATKINSON (Senior Member, IEEE) received the B.Eng. degree (Hons.) in electronic and electrical engineering, the M.Sc. degree in communications, control, and digital signal processing, and the Ph.D. degree in mobile communications systems from the University of Strathclyde, Glasgow, U.K., in 1993, 1995, and 2003, respectively. He is currently a Senior Lecturer with the institution. His research interests include data engineering and the application of machine learning algorithms to industrial problems, including cyber-security.

XAVIER BELLEKENS (Member, IEEE) received the bachelor's degree from Henam, Belgium, in 2010, the master's degree in ethical hacking and computer security from the University of Abertay, Dundee, in 2012, and the Ph.D. degree in electronic and electrical engineering from the University of Strathclyde, Glasgow, in 2016. He is currently a Chancellor's Fellow Lecturer with the Department of Electronic and Electrical Engineering, University of Strathclyde, where he has been working on cyber-security for critical infrastructures. Previously, he was a Lecturer in security and privacy with the Department of Cyber-Security, University of Abertay, where he led the Machine Learning for Cyber-Security Research Group. His current research interests include machine learning for cyber-security, autonomous distributed networks, the Internet of Things, and critical infrastructure protection.